Gene-expression profiles predict survival of patients with lung adenocarcinoma

David G. Beer[1], Sharon L.R. Kardia[2], Chiang-Ching Huang[3], Thomas J. Giordano[4], Albert M. Levin[2], David E. Misek[5], Lin Lin[3], Guoan Chen[1], Tarek G. Gharib[1], Dafydd G. Thomas[4], Michelle L. Lizyness[4], Rork Kuick[5], Satoru Hayasaka[3], Jeremy M.G. Taylor[3], Mark D. Iannettoni[1], Mark B. Orringer[1] & Samir Hanash[5]

Departments of [1]Surgery, [2]Epidemiology, [3]Biostatistics, [4]Pathology and [5]Pediatrics, University of Michigan, Ann Arbor, Michigan, USA
Correspondence should be addressed to D.G.B.; email: dgbeer@umich.edu.

Published online: 15 July 2002, doi:10.1038/nm733

Histopathology is insufficient to predict disease progression and clinical outcome in lung adenocarcinoma. Here we show that gene-expression profiles based on microarray analysis can be used to predict patient survival in early-stage lung adenocarcinomas. Genes most related to survival were identified with univariate Cox analysis. Using either two equivalent but independent training and testing sets, or 'leave-one-out' cross-validation analysis with all tumors, a risk index based on the top 50 genes identified low-risk and high-risk stage I lung adenocarcinomas, which differed significantly with respect to survival. This risk index was then validated using an independent sample of lung adenocarcinomas that predicted high- and low-risk groups. This index included genes not previously associated with survival. The identification of a set of genes that predict survival in early-stage lung adenocarcinoma allows delineation of a high-risk group that may benefit from adjuvant therapy.

Lung cancer remains the leading cause of cancer death in industrialized countries. Most patients with non-small cell lung cancer (NSCLC) present with advanced disease, and despite recent advances in multi-modality therapy, the overall 10-year survival rate remains a dismal 8–10%[1]. However, a significant minority of patients (~25–30%) with NSCLC have stage I disease and receive surgical intervention alone. Although 35–50% of patients with stage I disease will relapse within 5 years[2–4], it is not currently possible to identify specific high-risk patients.

Adenocarcinoma is currently the predominant histological subtype of NSCLC (refs. 1,5,6). Although morphological assessment of lung carcinomas can roughly stratify patients, there is a need to identify patients at high risk for recurrent or metastatic disease. Preoperative variables that affect survival of patients with NSCLC have been identified[7–10]. Tumor size, vascular invasion, poor differentiation, high tumor-proliferative index and several genetic alterations, including K-*ras* (refs. 11,12) and p53 (refs. 10,13) mutations, have prognostic significance. Multiple independently assessed genes or gene products have also been investigated to better predict patient prognosis in lung cancer[14–18]. Technologies that simultaneously analyze the expression of thousands of genes[19] can be used to correlate gene-expression patterns with numerous clinical parameters—including patient outcome—to better predict tumor behavior in individual patients[20]. Analyses of lung cancers using array technologies have identified subgroups of tumors that differ according to tumor type and histological subclasses and, to a lesser extent, survival among adenocarcinoma patients[21,22]. Here we correlated gene-expression profiles with clinical outcome in a cohort of patients with lung adenocarcinoma and identified specific genes that predict survival among patients with stage I disease. For further validation, we also show that the risk index predicted survival in an independent cohort of stage I lung adenocarcinomas.

Hierarchical profile clustering yields three tumor subsets

Using oligonucleotide arrays, we generated gene-expression profiles for 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors, as well as 10 non-neoplastic lung samples. Selected sample replicates showed high correlation among coefficients and reliable reproducibility. We determined transcript abundance using a custom algorithm and the data set was trimmed of genes expressed at extremely low levels, that is; genes were excluded if the measure of their 75th percentile value was less than 100. Although potentially resulting in the loss of some information, trimming in this manner decreased the possibility that the clustering algorithm would be strongly influenced by genes with little or no expression in these samples. Hierarchical clustering with the resulting 4,966 genes yielded 3 clusters of tumors (Fig. 1). All 10 non-neoplastic samples clustered tightly together within Cluster 1 (data not shown). We examined the relationships between cluster and patient and tumor characteristics (Fig. 1 and Supplementary Figure A online). There were associations between cluster and stage ($P = 0.030$) and between cluster and differentiation ($P = 0.01$). Cluster 1 contained the greatest percentage (42.8%) of well differentiated tumors, followed by Cluster 2 (27%) and Cluster 3 (4.7%). Cluster 3 contained the highest percentage of both poorly differentiated (47.6%) and stage III tumors (42.8%), yet contained 3 (14.3%) moderately differentiated and 1 (5%) well differentiated stage I tumor. Notably, 11 stage I tumors were present in Cluster 3, sug-

gesting a common gene-expression profile for this subset of stage I and stage III tumors.

For patients with stage I and stage III tumors, the average ages were 68.1 and 64.5 years and the percentage of smokers was 88.9% and 89.5%, respectively. Marginally significant associations between cluster and smoking history were observed ($P = 0.06$). A significant relationship between histopathological classification and cluster was only discernable for bronchioloalveolar adenocarcinomas (BAs), which were only present in Clusters 1 and 2 ($P = 0.0055$) and comprised 35.7% and 12.3% of tumors for Clusters 1 and 2, respectively.

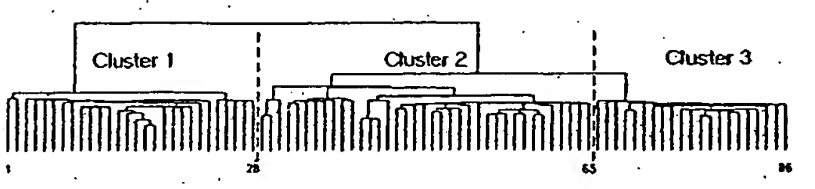


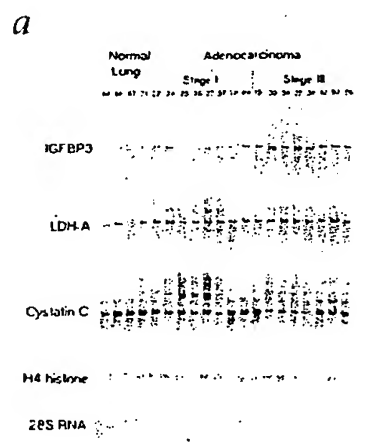
Fig. 1 Unsupervised classification analysis of lung adenocarcinomas. 3 classes of tumors identified by agglomerative hierarchical clustering of gene-expression profiles using the 4,966 expressed genes. Patient and histopathological information for each lung adenocarcinoma case by cluster designation and methods for K-*ras* 12/13th-codon mutational status and nuclear p53 protein accumulation are provided (Supplementary Figure A online). TN classification denotes information regarding patient tumor size and nodal involvement. Associations between cluster membership and patient or histopathological variables are indicated at significance level ($P \leq 0.05$).

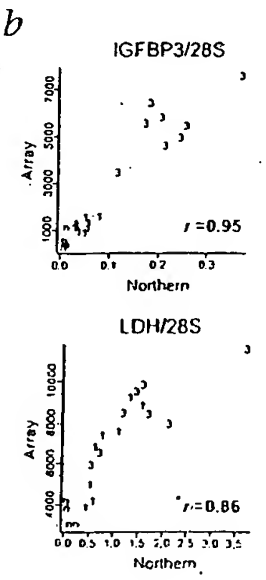
We examined the heterogeneity in gene-expression profiles based on the trimmed data set among normal lung samples and stage I and stage III adenocarcinomas by calculating correlation coefficients between all pairs of samples. In contrast to normal lung samples that displayed highly similar gene-expression profiles (median correlation, 0.9), both stage I and III lung tumors demonstrated much greater heterogeneity in their expression profiles with lower correlation coefficients (median values, 0.82 and 0.79, respectively).

Northern-blot and immunohistochemistry analyses

Of the 4,966 genes examined, 967 differed significantly between stage I and III adenocarcinomas, a number in excess of that expected by chance alone (248 at alpha level (α) = 0.05). Three genes were arbitrarily selected to verify the microarray expression data. The mRNA from 20 of the normal lung and tumor samples was examined by northern-blot hybridization with probes for insulin-like growth factor–binding protein 3 (*IGFBP3*), cystatin C and lactate dehydrogenase A (*LDH-A*) (Fig. 2*a*). Two gene probes not represented on the microarrays were used as controls, including histone H4, a potential index of overall cell proliferation, and 28S ribosomal RNA, a control for sample loading and transfer. The relative amounts of *IGFBP3*, cystatin C and *LDH-A* mRNA strongly correlated with microarray-based measurements (Fig. 2*b*). In both assays, *IGFBP3* and *LDH-A* mRNA levels increased from stage I to stage III adenocarcinomas and were higher than those in normal lung. Cystatin C mRNA levels were more variable but relatively greater in normal lung than tumors. These results suggest that the oligonucleotide microarrays provided reliable measures of gene expression. The tumors showed slightly greater histone H4 expression than the normal lung, likely reflecting increased proliferation of tumor cells.

Immunohistochemistry was performed for IGFBP3, cystatin C and HSP-70 to determine whether mRNA overexpression was reflected by an increase of their corresponding proteins in tumors.





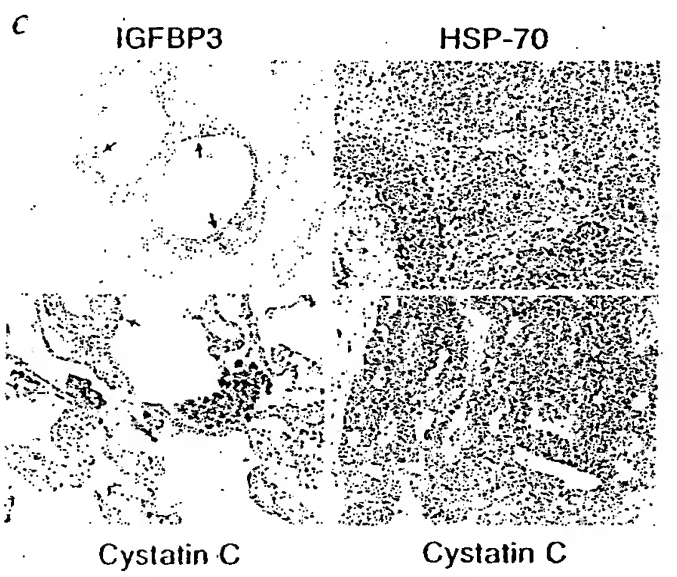


Fig. 2 Validation analyses of gene-expression profiling. *a*, Northern-blot analysis of selected candidate genes for verification of data obtained from oligonucleotide arrays. The same sample RNA for the 4 uninvolved lung, 8 stage I and 8 stage III tumors was used for the northern-blot and oligonucleotide array analyses. *b*, Correlation analysis of quantitative data obtained from oligonucleotide arrays and northern blots measured by integrated phosphorimager-based signals for the *IGFBP3* and *LDH-A* genes. The ratio of *IGFBP3*, cystatin C and *LDH-A* mRNA to *28S* rRNA was determined. The relative values for each gene from each sample are shown. n, non-neoplastic normal lung; 1, stage I tumors; 3, stage III tumors. *c*, Immunohistochemical analysis of IGFBP-3, HSP-70 and cystatin C in lung and lung adenocarcinomas. Cytoplasmic IGFBP-3 immunoreactivity in a neoplastic gland (tumor L22) with prominent apical staining (blue reactant staining, arrow, upper left). Diffuse cytoplasmic HSP-70 immunoreactivity (tumor L27), yet stromal elements show no reactivity (upper right). Normal lung parenchyma (lower left) shows cytoplasmic cystatin C immunoreactivity in alveolar pneumocytes (arrow) and intra-alveolar macrophages but tumor (L90) shows diffuse cytoplasmic cystatin C immunoreactivity with prominent apical staining (lower right). Magnification, ×200

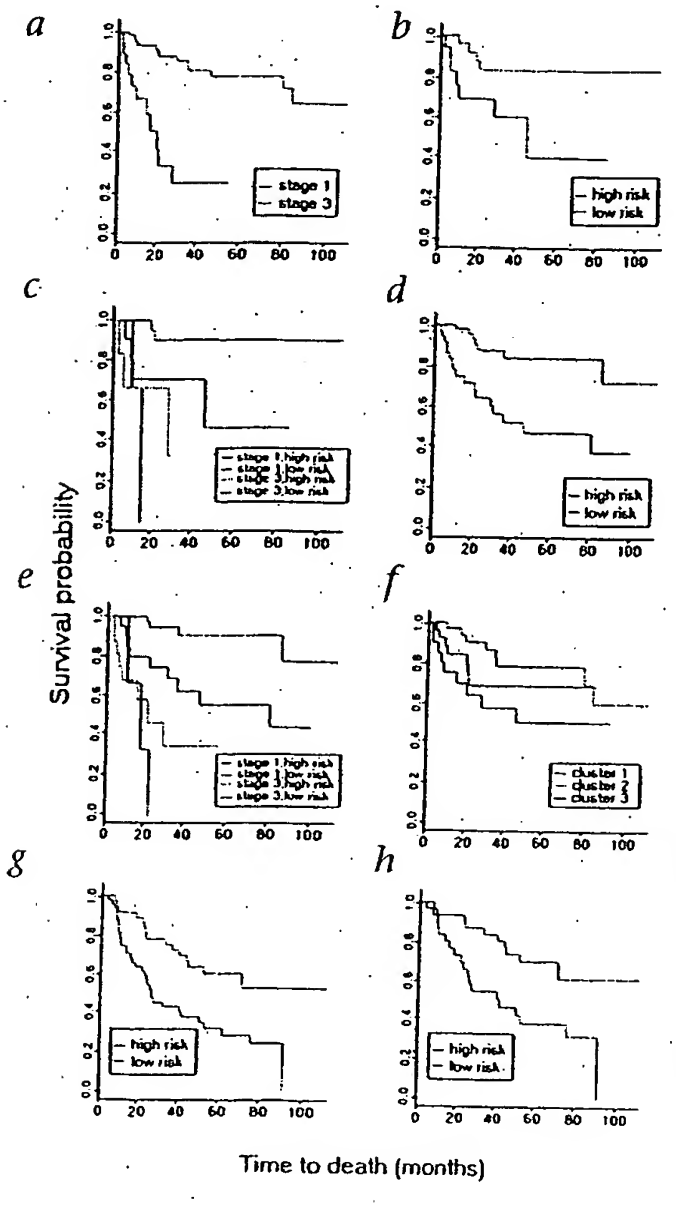


Fig. 3 Gene-expression profiles and patient survival. *a*, Relationship between tumor stage and patient survival (stage 1 and stage 3 differ significantly, $P < 0.0001$). *b*, Relationship between the survival in the 43 test samples and their risk assignments based on the 50-gene risk index estimated in the 43 training samples. The high- and low-risk groups differ significantly ($P = 0.024$). *c*, Relationship between patient survival and the risk assignments in test samples (in *b*) conditional for tumor stage. The high- and low-risk stage I groups differ significantly ($P = 0.028$), whereas stage III low- and high-risk groups did not ($P = 0.634$). *d*, Relationship between survival in the test cases and their risk assignments based on the 86 'leave-one-out' cross-validation of the 50-gene risk index. The high- and low-risk groups differ significantly ($P = 0.0006$). *e*, Relationship between test case's risk assignment and survival (in *d*) conditional on tumor stage. The high- and low-risk stage I lung adenocarcinoma groups differ significantly from each other ($P = 0.003$), whereas low- and high-risk stage III tumors do not. *f*, Relationship between tumor class identified by hierarchical clustering and patient survival. Survival for patients in Cluster 3 differed relative to the tumors in Cluster 2 ($P = 0.037$) and approached significance for Cluster 1 and 2 combined ($P = 0.06$). *g*, Analysis of the Michigan-based risk index using top cross-validated survival genes identify a low- and high-risk group in an independent cohort of 84 Massachusetts-based lung adenocarcinomas that are significantly different ($P = 0.003$). *h*, Among the 62 stage I lung adenocarcinomas in the Massachusetts sample, the high- and low-risk groups differed significantly ($P = 0.006$).

Immunoreactivity for both *IGFBP-3* and *HSP-70* (Fig. 2*c*) was detected in the cytoplasm of the adenocarcinomas, with little detectable reactivity in the stromal or inflammatory cells. Cystatin C was detected in alveolar pneumocytes and intra-alveolar macrophages in non-neoplastic lung parenchyma and also consistently in the cytoplasm of neoplastic cells.

Gene-expression profiles predict survival

As expected, Kaplan–Meier survival curves (Fig. 3*a*) and log-rank tests indicated poorer survival among stage III compared with stage I adenocarcinomas ($P = <0.0001$). Two statistical approaches were used to determine whether gene-expression profiles could predict survival using the data set of 4,966 genes. In one approach, equal numbers of randomly assigned stage I and stage III tumors constituted training ($n = 43$) and testing ($n = 43$) sets. In the training set, the top 10, 20, 50 or 75 genes were used to create risk indices that were evaluated for their association with survival using the 50th, 60th or 70th percentile cutoff points to categorize patients into high or low groups. The results were similar across cutoff points but the 50-gene risk index had the best overall association with survival in the training set. After conservatively choosing the 60th percentile cutoff point from the training set, we then applied this risk index and cutoff point to the testing set. The risk index of the top 50 genes correctly identified low- and high-risk individuals within the independent testing set ($P = 0.024$) (Fig. 3*b* and Supplementary Methods online). Notably, 11 stage I tumors were included in the high-risk subgroup. When this risk assignment was then conditionally examined for stage progression (Fig. 3*c*), low- and high-risk groups among stage I tumors were found to differ ($P = 0.028$) in their survival.

Identification of a robust set of survival genes

Although predictive of patient survival, a single training-testing set may not provide the most robust set of genes due to random sampling issues. Therefore, a 'leave-one-out' cross-validation approach was used to identify genes associated with survival from all 86-tumor samples. We first developed a 50-gene risk index in each training set, and then applied the risk index to the test case held out from the full set of tumors and assigned the held out tumor to the high- or low-risk groups (Fig. 3*d*). The high and low-risk subgroups determined in the test cases differed significantly in their overall survival ($P = 0.0006$). Among the larger group of stage I lung adenocarcinomas, the low-risk ($n = 46$) and high-risk ($n = 21$) groups had markedly different survival ($P = 0.003$) (Fig. 3*e*). Table 1 lists selected examples of the cumulative top 100 genes derived from this cross-validation procedure (complete list in Supplementary Table A online).

It was also noted that many of the stage I patients in the high-risk subgroup (Fig. 3*e*) were present in Cluster 3 (Fig. 1). Kaplan–Meier analysis (Fig. 3*f*) demonstrated a significantly worse survival ($P = 0.037$) for patients in Cluster 3 relative to patients in Cluster 2 and approaching significance for Cluster 1 and 2 combined ($P = 0.06$). This further indicates the important relationship between gene-expression profiles and patient survival, independent of disease stage.

Consistent with previous analyses of lung adenocarcinomas[23], 40% of stage I and 57.8% of stage III tumors had 12th or 13th codon K-*ras* gene mutations. Those patients with tumors containing K-*ras* mutations showed a trend of poorer survival, but

Table 1 Selected examples of the top 100 genes from cross-validation

Gene name	*P* (normal versus tumor t-test)	% Change in tumor	*P* (stage I versus stage III t-test)	% Change in stage III	Coefficient β	Unigene comment
						Apoptosis-related
CASP4	0.56	−6%	0.02	57%	0.0022	Caspase 4, apoptosis-related cysteine protease
P63	9.73E-04	37%	0.03	43%	0.0010	Transmembrane protein (63 kD), endoplasmic reticulum/Golgi intermediate compartment
						Cell adhesion and structure
KRT7	8.02E-08	126%	0.11	55%	0.0003	Keratin 7
LAMB1	0.14	−20%	0.01	60%	0.0027	Laminin, β1
						Cell cycle and growth regulators
BMP2	0.54	−21%	0.27	47%	0.0044	Bone morphogenetic protein 2
CDC6	1.31E-05	1070%	0.05	148%	0.0124	CDC6 (cell division cycle 6, *Saccharomyces cerevisiae* homolog)
S100P	2.10E-08	1572%	0.19	77%	0.0001	S100 calcium-binding protein P
SERPINE1	2.89E-03	72%	0.25	30%	0.0008	Serine (or cysteine) proteinase inhibitor, clade E (nexin)
STX1A	8.65E-08	54%	0.07	26%	0.0031	Syntaxin 1A (brain)
						Cell signaling
ADM	0.05	39%	0.04	117%	0.0016	adrenomedullin
AKAP 12	8.53E-03	-47%	0.05	214%	0.0010	A kinase (PRKA) anchor protein (gravin) 12
ARHE	0.06	−39%	0.05	87%	0.0092	ras homolog gene family, member E
GRB7	2.02E-03	38%	0.63	15%	0.0030	Growth factor receptor-bound protein 7
VEGF	6.50E-08	174%	0.02	85%	0.0013	Vascular endothelial growth factor
WNT10B	0.05	31%	0.48	20%	0.0022	Wingless-type MMTV integration site family, member 10B
						Chaperones
HSPA8	0.36	8%	9.01E-04	51%	0.0008	Heat-shock 70 kD protein 8
						Receptors
ERBB2	0.04	92%	0.37	120%	0.0013	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
FXYD3	0.10	111%	0.31	73%	0.0046	FXYD domain-containing ion transport regulator 3
SLC20A1	1.34E-03	58%	0.02	66%	0.0021	Solute carrier family 20 (phosphate transporter), member 1
						Enzymes, cellular metabolism
CSTB	1.57E-04	50%	0.15	34%	0.0001	Cystatin B (stefin B)
CTSL	0.48	−10%	0.03	67%	0.0007	Cathepsin L
CYP24	3.16E-06	N/A	0.97	2%	0.0008	Cytochrome P450, subfamily XXIV (vitamin D 24-hydroxylase)
FUT3	1.07E-07	114%	0.97	−1%	0.0033	Fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group included)
MLN64	0.20	32%	0.42	80%	0.0007	Steroidogenic acute regulatory protein related
PDE7A	0.12	33%	0.01	−35%	−0.0187	Phosphodiesterase 7A
PLGL	0.04	−68%	0.35	−170%	−0.0011	Plasminogen-like
SLC1A6	0.07	−32%	0.12	86%	0.0069	Solute carrier family 1 (high-affinity aspartate/glutamate transporter), member 6
						Transcription and translation
COPEB	0.10	−33%	0.26	25%	0.0016	Core promoter element binding protein
CRK	0.10	32%	0.03	48%	0.0098	v-crk avian sarcoma virus CT10 oncogene homolog
RELA	0.26	−7%	0.01	20%	0.0034	v-rel avian reticuloendotheliosis viral oncogene homolog A
						Unknown function
KIAA0005	2.21E-04	40%	0.02	45%	0.0010	KIAA0005 gene product
MGB1	0.27	125%	0.33	459%	0.0018	Mammaglobin 1

Bolded genes were also significant for survival in 43 tumor training set (Fig. 3*b*).

Table 1 Selected examples of the cumulative top 100 genes identified using training-testing, cross-validation of all 86 lung tumor samples. The percent change, as well as the direction, for the average values of the 10 non-neoplastic lung to all tumors, and for the 67 stage I to the 19 stage III tumors are shown. A positive coefficient β value is indicative of a relationship of gene expression to a poorer patient outcome. The genes are listed in potential functional categories. Genes that were also present in the top 50 survival genes using the 43-tumor training set (Fig. 3*b*) are indicated in bold type. Complete listing of the gene probe sets and annotated gene and unigene identifiers can be found in the Supplementary Methods.

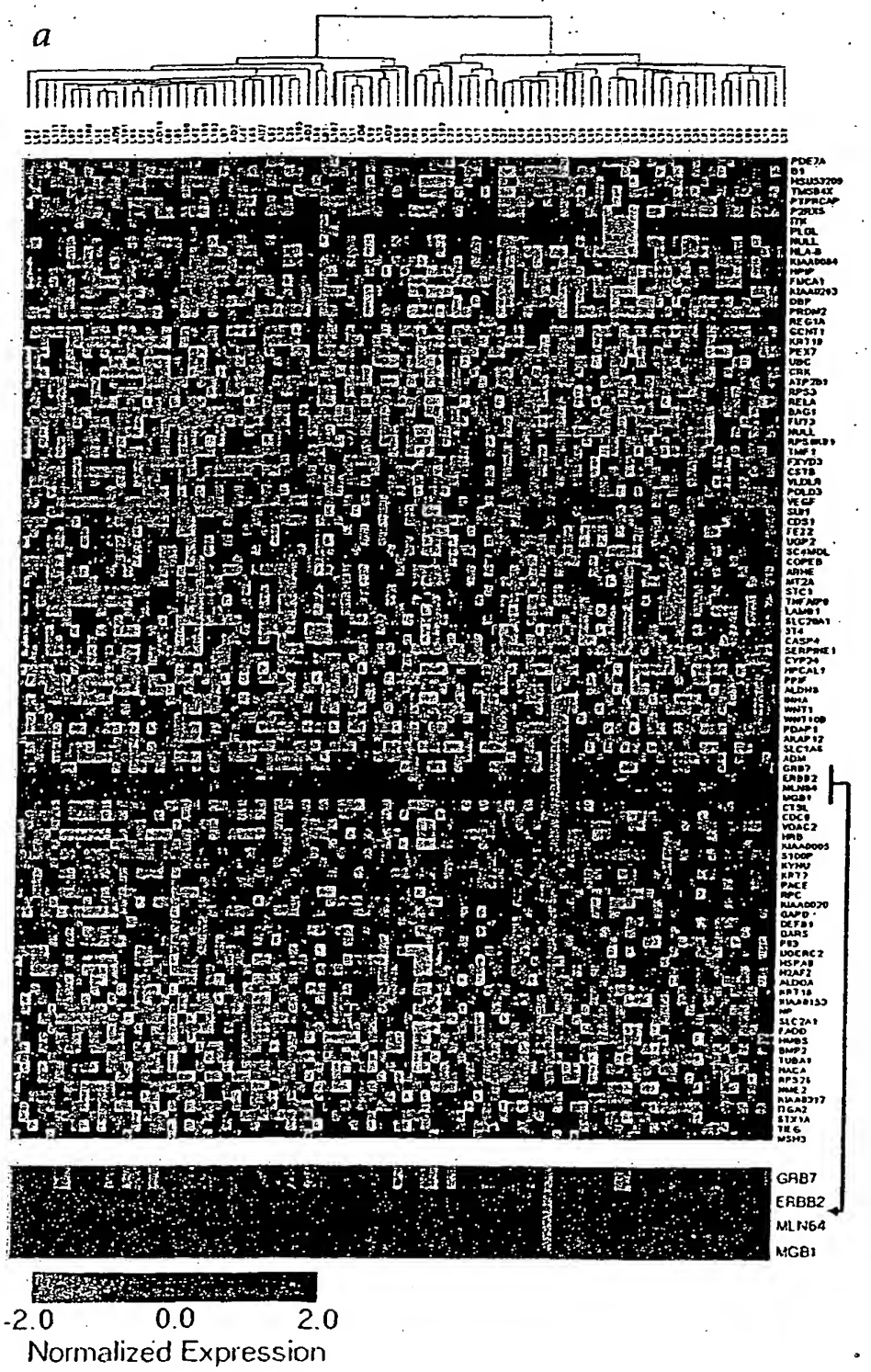


Fig. 4 Gene expression patterns of top survival genes *a*, Gene-expression patterns determined using agglomerative hierarchical clustering of the 86 lung adenocarcinomas against the 100 survival-related genes (Table 1) identified by the training-testing, cross-validation analysis. Substantially elevated (red) or decreased (green) expression of the genes is observed in individual tumors. Some tumors (black arrow and expanded area) show extremely elevated expression of specific genes. *b*, An outlier gene-expression pattern (>5 times the interquartile range among all samples) is observed for the *erbB2* and *Reg1A* genes (top left and right, respectively). The *S100P* and *crk* genes (bottom left and right, respectively) show a graded pattern of expression related to patient survival. ○, alive; ●, dead (also in *c*). *c*, The number of outliers per person identified in the top 100 genes plotted by survival distribution.

this difference did not reach statistical significance among all patients ($P = 0.25$), between patients within tumor clusters ($P = 0.41$) or when analyzed separately among stage I ($P = 0.22$) and stage III ($P = 0.53$) patients. Nuclear accumulation of p53 was detected in 17.9% stage I and in 22.2% stage III tumors. No significant relationship was observed for p53 staining and patient survival, cluster or tumor stage.

Confirmation using an independent set of adenocarcinomas

The robustness of our 50-gene risk index in predicting survival in lung adenocarcinomas was tested using oligonucleotide gene-expression data obtained from a completely independent (Massachusetts-based) sample of 84 lung adenocarcinomas (62 stage I, 14 stage II and 8 stage III; ref. 21, and dataset A at www.genome.wi.mit.edu/MPR/lung). To ensure equivalent power for testing and comparability of samples, the criteria for including tumors in the analysis were 40% or greater tumor cellularity, no mixed histology (that is, adenosquamous) and patient survival information. To obtain comparative gene-expression measures between the two data sets, gene sequences present on the U95A and HuGeneFL array were examined, and expression data for our top 50 cross-validation genes for all 84 Massachusetts samples were obtained and processed[24] (see also Supplementary Methods online). When we examined the risk assignment of these 84 samples, employing the identical cutoff point used for the 86 Michigan-based lung samples, we observed low- and high-risk groups (Fig. 3*g*; $P = 0.003$). Notably, among the 62 stage I tumors, high- and low-risk groups were observed that differed significantly ($P = 0.006$) in their survival (Fig. 3*h*).

Survival genes had graded and outlier expression patterns

A statistical and graphical analysis of the 100 survival-related genes (Table 1) clustered against all 86 tumors revealed individual tumors with substantially elevated expression in both a limited and larger number of genes (Fig. 4*a*). Among these genes, we observed two distinct patterns of expression related to patient survival. One pattern, designated 'outlier', included genes showing substantially elevated expression (greater than five times the interquartile range among all samples), whereas the other pattern, designated 'graded', was characterized by continuously distributed expression with patient survival (Fig. 4*b*). The *erb*B2 and *Reg*1A genes are examples of outlier expression patterns and *S100P* and *crk* genes of graded patterns. The number of outliers per person in the top 100 genes was identified and plotted according to survival times and events (Fig. 4*c*). Both stage I and stage III lung adenocarcinomas showed outlier gene patterns and 10 tumors contained 3 or more outlier genes.

Because gene amplification may result in increased gene expression, the nine genes with outlier expression patterns (*erb*B2, *SLC1A6*, *Wnt* 1, *MGB1*, *Reg*1A, *AKAP12*, *PACE*, *CYP24*, *KYNU*) and one gene with a graded expression pattern (*KRT18*) were examined using quantitative genomic PCR to evaluate genomic copy number (Fig. 5*a*). Gene amplification of *erb*B2 (17q12) was detected in tumor L94, which had the highest *erb*B2 mRNA expression (Fig. 4*a*). Gene amplification was not detected for any of the other seven tested genes in tumor L94, as well as in other tumors. The two genes most frequently demonstrating the outlier pattern in these lung adenocarcinomas were *KYNU* and *CYP24*, and were present in 10 and 9 tumors, respectively. *CYP24* has been described as a gene amplified and overexpressed in breast cancer[25], and these results indicate elevated expression in lung adenocarcinoma.

To determine whether the graded or outlier gene-expression patterns also occur at the protein-expression level, 10 of the 100

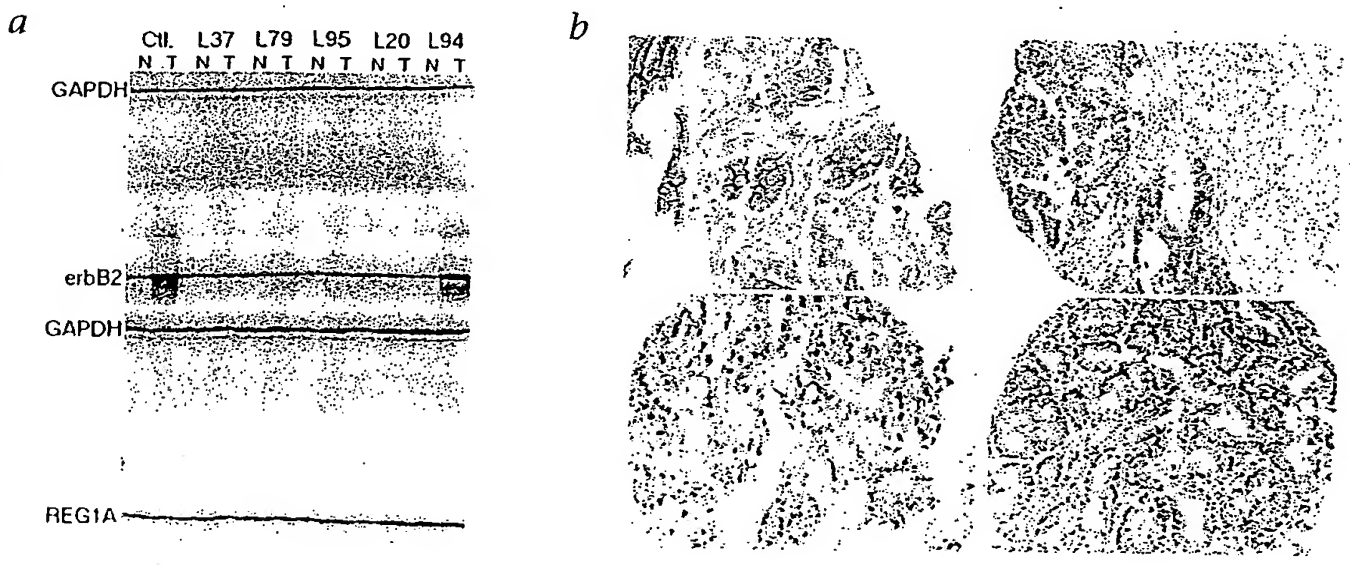


Fig. 5 Gene amplification and protein expression of survival-related genes. *a*, Analysis of potential gene amplification for 9 genes showing outlier expression patterns in the lung tumors (*erb*B2, *SLC1A6*, *Wnt* 1, *MGB1*, *Reg*1A, *AKAP12*, *PACE*, *CYP24* and *KYNU*) and examined using quantitative genomic PCR. A gene showing graded expression pattern (*KRT18*), and one gene (*PACE4*) with a similar chromosome location as *PACE*, were used as controls. Only *erb*B2 and *Reg*1A are shown. An esophageal adenocarcinoma with known high-level genomic amplification of *erb*B2 was used as a positive control and normal esophagus DNA was used as a negative control (Ctl). PCR fragments sizes were 343 bp for *GAPDH*, 166 bp for *erb*B2 and 126 bp for *Reg*1A. DNA is from normal lung (N) and tumor(T) from each patient (for example L37). *b*, Immunohistochemical analysis of survival related genes with lung adenocarcinoma microarrays using the tumors from this study. The transmembrane erbB2 protein (top left) expression is substantially increased in tumor L94 containing the amplified *erb*B2 gene (Fig. 4*a* and *b*). Expression of VEGF (top right) and S100P (bottom left) was located within the neoplastic cells and the pattern of immunoreactivity was consistent with the graded expression pattern demonstrated by their mRNA profiles. Expression of the oncogene *crk* (bottom right) was abundantly expressed in neoplastic lung cells. Magnification, ×400 (erbB2); ×200 (VEGF, S100P and crk).

top survival genes (Table 1) for which specific antibodies were available were chosen for immunohistochemical analysis using lung-tumor arrays from this study (Fig. 5*b*). Expression of membrane erbB2 protein was substantially increased in the *erb*B2-amplified tumor L94 and very low levels of expression were present in other tumors, consistent with mRNA-expression measurements (Fig. 4*a* and *b*). CDC6 protein expression was also substantially higher in tumor L94; consistent with mRNA levels (data not shown). Expression of vascular endothelial growth factor (VEGF) and S100P (Fig. 5*b*), as well as cytokeratin 18 (KRT18), cytokeratin 7 (KRT7) and fas-associated death domain (FADD) protein (data not shown), was located within the lung tumor cells and consistent with the graded expression pattern of the mRNA profiles. The oncogene *crk* showed both graded mRNA as well as a graded protein-expression pattern with survival, and was abundantly expressed in the tumor cells (Fig. 5*b*). These results indicate that many survival-associated genes are expressed at the protein level and demonstrate similar mRNA and protein-expression patterns.

Discussion

We used several approaches for the analysis of gene-expression data related to clinicopathological variables and patient survival. One approach, hierarchical clustering, was used to examine similarities among lung adenocarcinomas in their patterns of gene expression. Previous studies of lung tumors[21,22] have also used this method to describe subclasses of lung tumors. Here, we found three clusters that showed significant differences with respect to tumor stage and tumor differentiation. This suggests, as expected, that tumors with similar histological features of differentiation demonstrate similarities in gene expression. This feature also partly underlies the observed statistical association of tumor stage and cluster, as many of the higher-stage tumors, often poorly differentiated and previously associated with a reduced survival[9,10], were located in Cluster 3. Although this cluster contained the highest percentage of stage III tumors, it also contained a nearly equal mixture of stage I and stage III tumors and not all tumors were poorly differentiated. This indicates that a subset of stage I lung adenocarcinomas share gene-expression profiles with higher-stage tumors. Notably, 10 of the 11 stage I tumors found in Cluster 3 were the high-risk stage I tumors identified using the risk index in the 'leave-one-out' cross-validation.

In contrast to previous analyses of lung adenocarcinomas[21,22], we validated the expression data from the arrays. The strong correlation of northern-blot analysis and oligonucleotide-array data for gene expression in the same samples (Fig. 2*b*) indicates that these studies provide robust gene-expression estimates. Immunohistochemistry using the same tumor samples in tissue arrays demonstrates protein expression within the lung tumor cells. Together, these studies indicate that many of the genes identified using gene-expression profiles are likely relevant to lung adenocarcinoma. For example, *IGFBP3* gene expression is increased in lung adenocarcinomas (Fig. 2*c*). IGFBP3 protein modulates the autocrine or paracrine effects of insulin-like growth factors, elevated IGFBP3 expression is observed in colon cancer[26], and increased serum IGFBP3 is associated with progression in breast cancer[27]. Heat-shock protein 70 (HSP-70) is increased in lung adenocarcinomas of smokers[28] and is associated with increased metastatic potential in breast cancer[29]. Increased serum lactate dehydrogenase is correlated with tumor stage and tumor burden[30], and cystatin C, a cysteine protease inhibitor expressed in human lung cancers[31], is prognostic in some cancers[32]. The decreased expression of this protease inhibitor may affect the invasive properties of the tumor cell.

The cross-validation analytical strategy we used is particularly informative for these types of gene-expression analyses for disease outcome[33,34], and identification of cross-validated genes with a larger tumor cohort may help refine this risk index for use in a clinical setting. The gene-expression data also provide opportunities to observe overarching patterns that advance our understanding of associations between genes and disease. For example, the top 100 survival genes include those involved in signaling, cell cycle and growth, transcription, translation and metabolism. Expression of many of these genes is likely a function of increased proliferation and metabolism in the more aggressive tumors. Some genes, such as *erb*B2 and *Reg*1A (Fig. 4*a* and *b*), were highly overexpressed in a few patients having poor survival. In one tumor, the *erb*B2 gene was amplified (Fig. 5*a*), demonstrating that genomic changes may underlie the overexpression of a subset of these outlier genes. Immunohistochemistry confirmed protein overexpression in this patient's tumor (Fig. 5*b*). Notably, seven of the eight outlier genes were not amplified, indicating that other mechanisms underlie the increased mRNA expression of these survival-related genes.

Most genes showed a graded relationship between expression and patient survival. Genes such as that encoding VEGF, known to be strongly associated with survival in lung cancer[35,36] were identified as related to patient survival in our study. *VEGF* demonstrated a graded expression pattern, as did the S100P and *crk* oncogene (Fig. 5*b*). S100P is a calcium-regulated protein not previously reported in lung cancer. The *crk* gene, the cellular homolog of the v-*crk* oncogene, is a member of a family of adaptor proteins involved in signal transduction and interacts directly with c-jun N-terminal kinase 1 (JNK1)[37]. Although *crk* has not been shown to have a role lung cancer, its role in the MAP-kinase pathway, which leads to activation of matrix metalloproteinase secretion and cell invasion[38], indicates potential involvement in the the tumor cell invasion or metastasis of some lung adenocarcinomas. Among the many genes identified in this study, like *crk*, that may be causally involved in lung cancer progression (Table 1), some were related to survival in many patients, and others in only smaller subsets of patients. This result is consistent with the complex molecular architecture of tumors in general, the heterogeneity of lung adenocarcinomas in particular and the multiple mechanisms underlying tumor-cell survival, invasion and metastasis[39].

Our results demonstrate that a gene-expression risk profile—based on the genes most associated with patient survival—can distinguish stage I lung adenocarcinomas and differentiate prognoses. The particular genes that define the clusters, or are associated with survival, likely reflect the characteristics of the particular tumors included in the analysis. Current therapy for patients with stage I disease usually consists of surgical resection without adjuvant treatment[2,3]. Clearly, the identification of a high-risk group among patients with stage I disease would lead to consideration of additional therapeutic intervention for this group, possibly leading to improved survival of these patients.

Methods

Patient population. Sequential patients seen at the University of Michigan Hospital between May 1994 and July 2000 for stage I or stage III lung adenocarcinoma were evaluated for this study. Consent was received and the project was approved by the local Institutional Review Board. Primary tumors and adjacent non-neoplastic lung tissue were obtained at the time of

surgery. Peripheral portions of resected lung carcinomas were sectioned, evaluated by a study pathologist and compared with routine H&E sections of the same tumors, and utilized for mRNA isolation. Regions chosen for analysis contained a tumor cellularity greater than 70%, no mixed histology, potential metastatic origin, extensive lymphocytic infiltration or fibrosis. Tumors were histopathologically divided into two categories based on their growth pattern: bronchial-derived, if they exhibited invasive features with architectural destruction, and bronchioloalveolar, if they exhibited preservation of the lung architecture. All stage I patients received only surgical resection with intra-thoracic nodal sampling and no other treatments. Stage III patients received surgical resection plus chemotherapy and radiotherapy.

Gene-expression profiling and K-*ras* mutation analysis. RNA isolation, cRNA synthesis and gene-expression profiling were performed as described[24]. Details of gene annotation and K-*ras* mutation analysis are provided in supplementary information.

Northern-blot analysis. Total cellular RNA (10 μg) was separated in 1.2% agarose-formaldehyde gels and vacuum-transferred to Gene Screen Plus (NEN Life Science Products, Boston, Massachusetts). Hybridization conditions and probe labeling were as described[40]. Individual sequence-validated cDNA image clones for human *IGFBP3* (clone 1407750), *LDH-A* (clone 2420241), cystatin C (*CTS3*; clone 949938) were from Research Genetics (Huntsville, Alabama). The human histone H4 cDNA and the *28S* ribosomal RNA 26-mer oligonucleotide probe were prepared and labeled as described[40].

Gene-amplification analysis. 11 genes were selected for the analysis of genomic alterations. Primers were designed using PrimerSelect 4.05 Windows 32 software (DNASTAR, Madison, Wisconsin), avoiding pseudogenes or potential homologous regions. Forward and reverse primers for the genes are provided (Supplementary Methods online). Quantitative genomic-PCR was then applied and analyzed as described[41].

Immunohistochemical staining. The H&E-stained slides of all primary lung tumors were used to identify the most representative regions of each tumor and a tissue microarray (TMA) block was constructed as described[42]. Immunohistochemistry (IHC) was performed using both routine and sections from the TMA block as described[24]. Detailed methods and the concentrations used for all antibodies are provided in the Supplementary Methods.

Statistical methods. *t*-tests were used to identify differences in mean gene-expression levels between comparison groups. Agglomerative hierarchical clustering[43] was applied using the average linkage method to investigate whether there was evidence for natural groupings of tumor samples based on correlations between gene-expression profiles. To investigate the robustness of the clustering inference, gene-expression values were perturbed by adding random Gaussian error of magnitude obtained from a duplicate sample to each data point and then reclustered to determine concordance in the tumor's class membership. Pearson, χ^2 and Fisher's exact tests were used to assess whether cluster membership was associated with physical and genetic characteristics of the tumors.

To determine whether gene-expression profiles were associated with variability in survival times, 2 separate but complementary approaches were used. In the first approach, the 86 tumors were randomly assigned to equivalent training and testing sets consisting of equal numbers of stage I and III tumors in order to validate a novel risk-index function that captured the effect of many genes at once. In the second approach, cross-validation[44] was used to more robustly identify the genes associated with survival. Briefly, a 'leave-one-out' cross-validation procedure in which 85 of the 86 tumors (the training set) was used to identify genes that were univariately associated with survival. The risk index was defined as a linear combination of the gene-expression values for the top genes identified by univariate Cox proportional-hazard regression modeling[45], weighted by their estimated regression coefficients. Kaplan–Meier survival plots and log-rank tests were then used to assess whether the risk-index assignment to high/low categories was validated in the test set. A more detailed description is provided (Supplementary Methods online).

Note: Supplementary information is available on the Nature Medicine website.

Acknowledgments

We thank D. Sanders for technical assistance; D. Sing for assistance with the figures; and G. Omenn for critical reading of this manuscript. This work was supported by National Cancer Institute grant: U19 CA-85953 and the Tissue Core of the University of Michigan Comprehensive Cancer Center (NIH CA-46952).

Competing interests statement

The authors declare that they have no competing financial interests.

RECEIVED 5 APRIL; ACCEPTED 14 JUNE 2002

1. Fry, W.A., Phillips, J.L. & Menck, H.R. Ten-year survey of lung cancer treatments and survival in hospitals in the United States. *Cancer* 86, 1867–1876 (1999).
2. Williams, D.E. *et al.* Survival of patients surgically treated for stage I lung cancer. *J. Thorac. Cardiovasc. Surg.* 82, 70–76 (1981).
3. Pairolero, P.C. *et al.* Postsurgical stage I bronchogenic carcinoma: Morbid implications of recurrent disease. *Ann. Thorac. Surg.* 38, 331–338 (1984).
4. Naruke, T. *et al.* Prognosis and survival in resected carcinoma based on the new international staging system. *J. Thorac. Cardiovasc. Surg.* 96, 440–447 (1988).
5. Kaisermann, M.C. *et al.* Evolving features of lung adenocarcinoma in Rio de Janeiro, Brazil. *Oncol. Rep.* 8, 189–192 (2001).
6. Roggli, V.L. *et al.* Lung cancer heterogeneity: A blinded and randomized study of 100 consecutive cases. *Hum. Pathol.* 16, 569–579 (1985).
7. Gail, M.H. *et al.* Prognostic factors in patients with resected stage I non-small cell lung cancer: A report from the Lung Cancer Study Group. *Cancer* 54, 1802–1813 (1984).
8. Takise, A. *et al.* Histopathologic prognostic factors in adenocarcinomas of the peripheral lung less than 2 cm in diameter. *Cancer* 61, 2083–2088 (1988).
9. Ichinose, Y. *et al.* Is T factor of the TMN staging system a predominant prognostic factor in pathologic stage I non-small cell lung cancer. *J. Thorac. Cardiovasc. Surg.* 106, 90–94 (1993).
10. Harpole, D.H. *et al.* A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation, histopathology, and oncoprotein expression. *Cancer Res.* 55, 51–56 (1995).
11. Rodenhuis, S. *et al.* Mutational activation of the K-*ras* oncogene: A possible pathogenic factor in adenocarcinoma of the lung. *N. Engl. J. Med.* 317, 929–935 (1987).
12. Slebos, R.J.C. *et al.* K-*ras* oncogene activation as a prognostic marker in adenocarcinoma of the lung. *N. Engl. J. Med.* 323, 561–565 (1990).
13. Horio, Y. *et al.* Prognostic significance of p53 mutations and 3p deletions in primary resected non-small cell lung cancer. *Cancer Res.* 53, 1–4 (1993).
14. Kern, J.A. *et al.* C-*erbB*-2 expression and codon 12 K-*ras* mutations both predict shortened survival for patients with pulmonary adenocarcinomas. *J. Clin. Invest.* 93, 516–520 (1994).
15. Ebina, M. *et al.* Relationship of p53 overexpresson and up-regulation of proliferating cell nuclear antigen with the clinical course of non-small cell lung cancer. *Cancer Res.* 54, 2496–2503 (1994).
16. Mehdi, S.A. *et al.* Prognostic markers in resected stage I and II non-small cell lung cancer: an analysis of 260 patients with 5 year follow-up. *Clin. Lung Cancer* 1, 59–67 (1997).
17. Schneider, P.M. *et al.* Multiple molecular marker testing (p53, c-Ki-*ras*, c-erbB-2) improves estimation of prognosis in potentially curative resected non-small cell lung cancer. *Br. J. Cancer* 83, 473–479 (2000).
18. Herbst, R.S. *et al.* Differential expression of E-cadherin and type IV collagenase genes predicts outcome in patients with stage I non-small cell lung carcinoma. *Clin. Can. Res.* 6, 790–797 (2000).
19. Liotta, L. & Petricion, E. Molecular profiling of human cancer. *Nature Rev. Genet.* 1, 48–56 (2000).
20. Golub, T.R. Editorial: Genome-wide views of cancer. *N. Engl. J. Med.* 344, 601–602 (2001).
21. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* 98, 13790–13795 (2001).
22. Garber, M.E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* 98, 13784–13789 (2001).
23. Mills, N.E. *et al.* Increased prevalence of K-*ras* oncogene mutations in lung adenocarcinoma. *Cancer Res.* 55, 1444–1447 (1995).
24. Giordano T.J. *et al.* Organ-specific molecular classification of lung, colon and ovarian adenocarcinomas using gene expression profiles. *Am. J. Pathol.* 159, 1231–1238 (2001).
25. Albertson, D.G. *et al.* Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genet.* 25, 144–146 (2000).
26. Kansra, S. *et al.* IGFBP-3 mediates TGF β1 proliferative response in colon cancer cells. *Int. J. Cancer* 87, 373–378 (2000).
27. Vadgama J.V. *et al.* Plasma insulin-like growth factor-I and serum IGF-binding protein 3 can be associated with the progression of breast cancer, and predict the risk of recurrence and the probability of survival in African-American and Hispanic

women. *Oncology* 57, 330–340 (1999).
28. Volm, M., Mattern, J. & Stammler, G. Up-regulation of heat shock protein 70 in adenocarcinoma of the lung in smokers. *Anticancer Res.* 15, 2607–2609 (1995).
29. Ciocca, D.R. *et al.* Heat shock protein hsp70 in patients with auxillary lymph node-positive breast cancer:prognostic implications. *J. Natl. Cancer. Inst.* 85, 570–574 (1993).
30. Rotenberg, Z. *et al.* Total lactate dehydrogenase and its isoenzymes in serum of patients with non-small cell lung cancer. *Clin. Chem.* 34, 668–670 (1988).
31. Krepela, E. *et al.* Cysteine proteases and cysteine protease inhibitors in non-small cell lung cancer. *Neoplasma* 45, 318–331 (1998).
32. Kos, J. *et al.* Cysteine proteinases and their inhibitors in extracellular fluids: Markers for diagnosis and prognosis in cancer. *Int. J. Biol. Markers* 15, 84–89 (2000).
33. Golub, T.R. *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537 (1999).
34. Hedenfalk, I. *et al.* Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344, 539–548 (2001).
35. Ohta, Y. *et al.* Vascular endothelial growth factor and lymph node metastasis in primary lung cancer. *Br. J. Cancer.* 76, 1041–1045 (1997).
36. Shibusa, T., Shijubo, N. & Abe, S. Tumor angiogenesis and vascular endothelial growth factor expression in stage I lung adenocarcinoma. *Clin. Cancer Res.* 4, 1483–1487 (1998).
37. Girardin, S.E. & Yaniv, M. A direct interaction between JNK1 and CrkII is critical for Rac1-induced JNK activation. *EMBO J.* 20, 3437–3446 (2001).
38. Liu, E. *et al.* The Ras-mitogen-activated protein kinase pathway is critical for the activation of matrix metalloproteinase secretion and the invasiveness in v-crk-transformed 3Y1. *Cancer Res.* 60, 2361–64 (2000).
39. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* 100, 57–70 (2000).
40. Hanson, L.A. *et al.* Expression of the glucocorticoid receptor and K-*ras* genes in urethan-induced mouse lung tumors and transformed cell lines. *Exp. Lung. Res.* 17, 371–387 (1991).
41. Lin, L. *et al.* A minimal critical region of the 8p22-23 amplicon in esophageal adenocarcinomas defined using STS-amplification mapping and quantitative PCR includes the GATA-4 gene. *Cancer Res.* 60, 1341–1347 (2000).
42. Kononen, J. *et al.* Tissue microarrays for high throughput molecular profiling of tumor specimens. *Nature Med.* 4, 844–847 (1998).
43. Johnson, R. & Wichern, D.W. *Applied Multivariate Statistical Analysis.* 543–578 (Prentice Hall, New Jersey, 1988).
44. Stone, M. Asymptomics for and against cross-validation. *Biometrika* 64, 29–38 (1977).
45. Cox, D.R. Regression models and life tables. *J.R. Stat. Soc.* 34, 187–220 (1972).

BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 585–596

Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts

Dov Greenbaum[3,†], Ronald Jansen[1,†] and Mark Gerstein[1,2,*]

[1]Departments of Molecular Biophysics & Biochemistry, [2]Computer Science and [3]Genetics, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA

Received on July 2, 2001; revised on October 5, 2001; accepted on October 22, 2001

ABSTRACT

Motivation: Protein abundance is related to mRNA expression through many different cellular processes. Up to now, there have been conflicting results on how correlated the levels of these two quantities are. Given that expression and abundance data are significantly more complex and noisy than the underlying genomic sequence information, it is reasonable to simplify and average them in terms of broad proteomic categories and features (e.g. functions or secondary structures), for understanding their relationship. Furthermore, it will be essential to integrate, within a common framework, the results of many varied experiments by different investigators. This will allow one to survey the characteristics of highly expressed genes and proteins.

Results: To this end, we outline a formalism for merging and scaling many different gene expression and protein abundance data sets into a comprehensive reference set, and we develop an approach for analyzing this in terms of broad categories, such as composition, function, structure and localization. As the various experiments are not always done using the same set of genes, sampling bias becomes a central issue, and our formalism is designed to explicitly show this and correct for it. We apply our formalism to the currently available gene expression and protein abundance data for yeast. Overall, we found substantial agreement between gene expression and protein abundance, in terms of the enrichment of structural and functional categories. This agreement, which was considerably greater than the simple correlation between these quantities for individual genes, reflects the way broad categories collect many individual measurements into simple, robust averages. In particular, we found that in comparison to the population of genes in the yeast genome, the cellular populations of transcripts and proteins (weighted by their respective abundances, the transcriptome and what we dub the translatome) were both enriched in: (i) the small amino acids Val, Gly, and Ala; (ii) low molecular weight proteins; (iii) helices and sheets relative to coils; (iv) cytoplasmic proteins relative to nuclear ones; and (v) proteins involved in 'protein synthesis,' 'cell structure,' and 'energy production.'

Supplementary information: http://genecensus.org/expression/translatome

Contact: mark.gerstein@yale.edu

*To whom correspondence should be addressed.
†These authors contributed equally to this work.

INTRODUCTION

High throughput experimentation, measuring mRNA (Schena *et al.*, 1995; Eisen and Brown, 1999; Ferea and Brown, 1999; Lipshutz *et al.*, 1999) and protein expression (Anderson and Seilhamer, 1997; Futcher *et al.*, 1999; Gygi *et al.*, 1999a; Ross-Macdonald *et al.*, 1999; Lopez, 2000; MacBeath and Schreiber, 2000; Nelson *et al.*, 2000; Zhu *et al.*, 2000) are currently the single richest source of genomic information. However, how to best interpret this data is still an open question (Bassett *et al.*, 1996; Wittes and Friedman, 1999; Zhang, 1999; Gerstein and Jansen, 2000; Searls, 2000; Sherlock, 2000; Claverie, 1999; Einarson and Golemis, 2000; Epstein and Butow, 2000; Shapiro and Harris, 2000). Understanding how protein abundance is related to mRNA transcript levels is essential for interpreting gene expression, protein interactions, structures and functions in a cellular system (Hatzimanikatis *et al.*, 1999). Moreover, as protein concentration is the more relevant variable with respect to enzyme activity, it connects genomics to the physical chemistry of the cell (Kidd *et al.*, 2001). Protein abundance may also be invaluable for diagnostics and for determining drug targets (Corthals *et al.*, 2000).

Previously, we surveyed the population of protein features—such as folds, amino acid composition, and functions—in yeast, and other recently sequenced genomes (Gerstein, 1997, 1998a,b; Gerstein and Hegyi, 1998; Hegyi and Gerstein, 1999; Das and Gerstein, 2000; Lin and Gerstein, 2000), and we extended this concept to compare the population of features in the yeast transcriptome to that in the genome (Drawid *et al.*, 2000; Jansen and Gerstein, 2000). Others have also done related work (Frishman and Mewes, 1997; Tatusov *et al.*, 1997; Jones, 1998; Wallin and von Heijne, 1998; Frishman and Mewes, 1999; Wolf *et al.*, 1999). Here, we present a new methodology to compare the features of the mRNA expression population with the protein abundance population.

Precise terminology is essential for this comparison. Unfortunately, 'proteome' is used inconsistently. Proteome can logically be used to describe all the distinct proteins in the genome (Qi *et al.*, 1996; Cavalcoli *et al.*, 1997; Fey *et al.*, 1997; Garrels *et al.*, 1997; Gaasterland, 1999; Jones, 1999; Sali, 1999; Tekaia *et al.*, 1999; Bairoch, 2000; Cambillau and Claverie, 2000; Doolittle, 2000; Pandey and Mann, 2000; Rubin *et al.*, 2000) and, in this context, it is equivalent to what others may refer to as the coding part of the genome. However, in papers on two-dimensional (2D) electrophoresis, it is often used to describe the sum total of proteins in a cell, taking into account the different levels of protein abundance (Shevchenko *et al.*, 1996; Gygi *et al.*, 2000a; Lopez, 2000; Washburn and Yates, 2000). In an effort to be clear, we propose the term 'translatome' for this second usage of proteome.

With this definition, we are able to refer compactly to three different cellular populations. These are illustrated in Figure 1.

(i) We use the term *genome* when we refer to the population of open reading frames, where each ORF counts once.

(ii) We use the term *transcriptome* when we refer to the population of mRNA transcripts. This term was originally coined by Velculescu *et al.* (1997). Note that each ORF may give rise to different numbers of transcripts. Consequently, the transcriptome is essentially the same as the genome but with each ORF weighted by its expression level.

(iii) The next level is the cellular population of proteins. As each protein represents a translated transcript, we make an analogy with the term transcriptome and use the term *translatome* as described above to describe this third population. Thus, the translatome is a subset of the genome where each ORF is weighted by its associated level of protein abundance.

Note that one could also less compactly call the translatome a 'weighted proteome.' However, doing so assumes one of the two aforementioned definitions of proteome. To avoid ambiguity, we studiously avoid the use of proteome altogether in the paper.

Differences between the translatome and the transcriptome exist given that transcripts from different genes can give rise to different numbers of proteins, due to different rates of translation and protein degradation. Post-transcriptional modifications further affect the translatome.

In our analysis of the transcriptome and translatome, we focus on global protein features rather than the comparison of individual genes. Previous analyses have shown that differences between mRNA expression and protein abundance levels can be quite dramatic for individual genes. This may either be due to the noise in the data or to fundamental biological processes. However, our analyses show that the variation between transcriptome and translatome is much smaller for global properties that are computed by averaging over the properties of many individual genes.

METHODS

Data sources used

For our analysis we culled many divergent data sets, representing protein abundance and mRNA expression experiments and also other sources of genome annotation. These are all summarized in Table 1.

Biases in the data

The databases that annotate the specific genes may not always be accurate (Ishii *et al.*, 2000). Gene Chip experiments suffer with regard to cross hybridization and the saturation of probes. SAGE data degrades for lowly expressed mRNAs. 2D gels are unable to resolve membrane proteins (approximately 30% of the genome) and basic proteins (Gerstein, 1998c; Krogh *et al.*, 2001). In addition, the procedures for identification and quantification of the protein spots are subject to uncertainties (Haynes and Yates, 2000). Human biases include the lack of low abundance proteins (Fey and Larsen, 2001; Gygi *et al.*, 2000b; Harry *et al.*, 2000) and the differences between laboratories in sample preparation. Our reference expression data set attempts to resolve these problems.

Data set scaling

A reference set for mRNA expression. With many different mRNA expression data sets available, it is worthwhile to integrate them into a single unified reference set, with the intention of reducing the noise and errors contained in the individual data sets and to obtain a unified estimate of the normal expression state in a cell.

We adopt an iterative scaling and merging formalism,

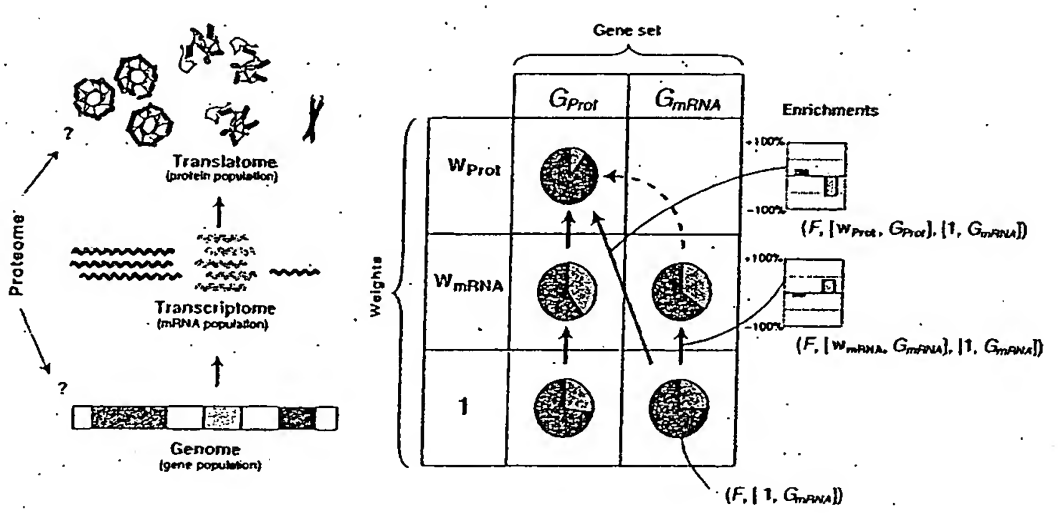


Fig. 1. Schematic overview of the analysis. On the left-side we outline the terms we use to describe the process of gene expression. The coding section of the genome is transcribed into a population of mRNA transcripts called the 'transcriptome.' The transcripts in turn are translated to a population of proteins; we use the term 'translatome' for this protein population rather than the alternative 'proteome' because the latter term may be confounded with the protein complement of the genome (which is not necessarily associated with a quantitative abundance level).

The matrix in the middle schematically shows an analysis of the three stages of expression. In general, we define a protein 'population' as a set of genes associated with a corresponding number of expression or abundance levels ('weights'). In the matrix each row represents a weight and each column a gene set. In particular, we differentiate between the mRNA reference expression set ($G_{mRNA} = G_{Gen}$), which essentially covers the complete genome, and the reference protein abundance set (G_{Prot}) which contains the proteins in data sets 2-DE #1 and 2-DE #2 (see Table 1) because the protein abundance set is a significantly smaller subset of the genome. By definition, this subset contains only proteins that can be identified by 2-D gel electrophoresis and is therefore biased in this sense. The enrichment figures throughout this paper, through a comparison of the right- and left-sides of this figure, show the results of the experimental biases of 2D gels on the data set. Each pie chart represents a composition of a particular protein feature F (for instance, an amino acid composition) in a population (represented by the symbol μ). We can further look at the 'enrichment' of this feature in one population relative to another (represented by the symbol Δ, see Section 'Methods' for an explanation of the formalism).

which we summarize below. We present a more detailed review of the methods on our web site.

We start with the values of one gene chip data set U_i where i is used throughout as a subscript to denote gene number. We then transform the values of the next Gene Chip data set X_i to Y_i with the following non-linear regression: $\min \sum_i (Y_i - U_i)^2$ with $Y_i = AX_i^B$ where A and B are the parameters of the regression. Note that two Gene Chip sets may not be defined for the same set of genes, so we have to perform the fit only over the genes common to both sets. The motivation for scaling is that the dynamic range of observed expression levels varies somewhat between different data sets, although cell types and growth conditions are very similar. Reasons for disparity may include different calibration procedures for relating fluorescence intensity to a cellular concentration (measured in copies of transcripts per cell) or different protocols for harvesting and reverse-transcribing the cellular mRNA.

We then merge and average the data to create a new reference set V as follows:

$$\text{If } U_i \text{ and } Y_i \text{ are both defined for gene } i \text{ and } \frac{|Y_i - U_i|}{Y_i + U_i} < \alpha$$

Then $V_i = \frac{1}{2}(Y_i + U_i)$

Else if only Y_i exists, $V_i = Y_i$

Else $V_i = U_i$.

As presented above, where only one data set has a value for the corresponding ORF, we incorporated that value and did not exclude it. When both data sets have values for an ORF, we averaged the values if they were within 15% of each other; otherwise, we just stayed with the original chip data set U_i. We used $\alpha = 15\%$ in order to prevent outliers from skewing the result. This 15% value is a reasonable threshold for excluding outliers though other values (e.g. 10 or 20%) would give similar results (data not shown). Other data sets are subsequently included in the same procedure, continuing the iteration from the new

expression values V_i. The initial iteration starts with the Young Expression Set, as U_i, since we have the highest confidence in its accuracy.

The SAGE data (Velculescu *et al.*, 1997) was not included in the above procedure since it is of a fundamentally different nature. An advantage of the SAGE technology over Gene Chips is that there is no possible signal saturation for high expression levels, as is possible for chips (Futcher *et al.*, 1999). Conversely, SAGE values are less reliable for lowly expressed genes since there is a chance that one might not sequence a SAGE tag corresponding to such a gene altogether. Therefore, if after the last iteration, the average Gene Chip expression level V_i was both above a certain threshold β and below the SAGE expression level S_i for the same gene, it was replaced with the SAGE value; otherwise the average Gene Chip value was kept. This gave us our final expression set w_{mRNA}. Our treatment of the SAGE data is modeled after that in Futcher *et al.* (1999), and like them, we used $\beta = 16$.

This incorporation of the SAGE data into the reference data set ensures that the highly expressed outliers are as accurate as possible.

Rather than plain arithmetic averaging, this overall scaling procedure with the α cutoff avoids 'artificial averages' that combine very different values for a particular gene. Some expression values might be statistical outliers. In addition, it may be possible that the expression levels of a variety of genes can only be within mutually exclusive ranges or modes, such as when two alternative pathways are switched on or off. Simply averaging these would give values that are less representative of the particular mode values. This situation is analogous to that in averaging together an ensemble of protein structures (i.e from NMR structure determination). Each structure could be stereochemically correct, with all side-chain atoms in predefined rotamer configurations. However, an average of all structures could yield one that is stereochemically incorrect if this involved averaging over particular side-chains in different rotameric states.

With regard to our regression analysis, we have investigated both non-linear and linear fits but found a non-linear procedure to be more advantageous. The non-linear relationship between different expression data sets perhaps reflects saturation in one or more of the Gene Chips—not an uncommon phenomenon. This non-linearity is immediately evident on scatter plots of two data sets against one another (see website). Accordingly, the non-linear fit produces a smaller residual than the linear fit: 98 297 (non-linear) versus 122 182 (linear) for the scaling of the Church data set and 59 828 (non-linear) versus 67 462 (linear) for the Samson data set.

A reference set for protein abundance. We followed a similar procedure to calculate a reference protein abundance set from the two gel electrophoresis data sets. We first scaled the two data sets against the mRNA expression reference data set, getting regression parameters C_j and D_j:

$$\min \sum_i (P_{i,j} - C_j w_{\mathrm{mRNA},i}^{D_j})^2$$

where the subscript j indicates the data set 2-DE #1 or 2-DE #2 respectively; $P_{i,j}$ is the protein abundance value in data set j, and $w_{\mathrm{mRNA},i}$ the corresponding reference expression value, and C_j and D_j are the parameters of the non-linear regression.

Using these parameters, we transformed the values of set 2-DE #2 onto 2-DE #1. Then we combined both sets into the reference protein set w_{Prot} by averaging them, if both values existed. Otherwise, by using the existing value, viz:

$$Q_{i,2} \equiv C_1 \left(\frac{P_{i,2}}{C_2} \right)^{D_1/D_2}$$

$w_{\mathrm{Prot},i} = (P_{i,1} + Q_{i,2})/2$ if both $P_{i,1}$ and $Q_{i,2}$ exist.
Else if only $P_{i,1}$ exists, $w_{\mathrm{Prot},i} = P_{i,1}$
Else if $Q_{i,2}$ exists, $w_{\mathrm{Prot},i} = Q_{i,2}$.

Enrichment of features

Formalism. In the next part of our analysis, we want to group a number of proteins together into various categories based on common features and characterize those features that are enriched in one population relative to another, i.e. the translatome population of proteins as measured by 2D gels relative to the transcriptome population of transcripts or the genome population of genes. To this end, we set up a formalism that could be applied universally to all the attributes that we were interested in. Due to the limitations of the experiments, the translatome, transcriptome, and genome populations are defined on different sets of genes, and sometimes we want to remove this 'selection bias' by forcing them to be compared on exactly the same set of genes. This is a key aspect of our formalism as presented in Figure 1.

We call an entity like $[w, G]$ a 'population,' where G is a set describing a particular selection of genes from the genome and w is vector of weights associated with each element of this population. In particular, we focus on three main populations here:

(i) $[1, G_{\mathrm{Gen}}]$ is the population of genes in the genome, all 6280 genes weighted once ($w = 1$);

(ii) $[w_{\mathrm{mRNA}}, G_{\mathrm{mRNA}}]$ is the observed population of the transcripts in the transcriptome, i.e. the 6249 genes in the reference expression set weighted by their reference expression value;

(iii) $[w_{Prot}, G_{Prot}]$ is the observed cellular population of the proteins in the translatome, i.e. the 181 genes in the reference abundance set weighted by their reference abundance value.

(The set of genes in the genome G_{Gen} is approximately equal to the genes in set G_{mRNA}, such that we can use both symbols interchangeably.) We can also use this notation to describe specific experiments—e.g. $[w_{lacZ}, G_{lacZ}]$ describes the gene set and weights relating to the transposon abundance set.

Furthermore, we define F_j as the value of a feature F in ORF j. For example, F could be the composition of leucine (a real number) or a binary value (0 or 1) indicating whether an ORF contains a trans-membrane segment. Given these definitions, the weighted average of feature F in population $[w, G]$ is:

$$\mu(F, [w, G]) \equiv \frac{\sum_{j \in G} w_j F_j}{\sum_{j \in G} w_j}.$$

The weighted averages of two populations $[w, G]$ and $[v, S]$ can be compared by simply looking at their relative difference Δ:

$$\Delta(F, [v, S], [w, G]) = \frac{\mu(F, [v, S]) - \mu(F, [w, G])}{\mu(F, [w, G])}$$

where v and w are weights for the sets of ORFs S and G respectively. We call Δ the 'enrichment' of feature F because it indicates whether F is enriched (if Δ is positive) or depleted (if Δ is negative) in population $[v, S]$ relative to $[w, G]$.

Usually, the gene set G is defined by the particular experiment, for which the weight w was measured. However, it is also possible to combine the gene set associated with one experiment with expression levels from another set. One may want to do this to compute the enrichment only on the genes common to both populations, for which there are defined values for both w and v, viz: $\Delta(F, [v, S \cap G], [w, S \cap G])$. In practice, this is most relevant for comparing G_{Prot} and G_{mRNA}. Since G_{Prot} is completely a subset of G_{mRNA}, we need not explicitly deal with intersections if we calculate all statistics directly over G_{Prot}.

One can adjust the weight vectors to take into account different types of averaging. For instance, when computing the amino acid composition ($F = aa$) from the amino acid compositions of individual ORFs $F_j = aa_j$ ($\forall j \in G$), we weight by ORF length. In the case of expression weights, we have:

$$w_j = N_j w_{mRNA,j} \quad \forall j \in G$$

where N_j is a measure of the length of ORF j (such as the number of amino acids).

On the other hand, when computing the average molecular weight per amino acid, we need to normalize by the number of amino acids per ORF, which is equivalent to choosing the following weights:

$$w_j = \frac{w_{mRNA,j}}{N_j} \quad \forall j \in G.$$

Application of methodology to quantitative abundance sets

Having defined our formalism, we applied it to a diverse set of protein features in yeast.

Amino acid enrichment. As shown in Figure 2a, we used our methodology to measure the enrichment of individual amino acids in both the translatome and the transcriptome relative to the genome. We found that three amino acids—valine, glycine and alanine—were consistently enriched in both transcriptome and translatome populations.

In Figure 2a we compare different gene sets. In Figure 2b we focus mainly on the variation in enrichments when all the comparisons are restricted to the set of 181 genes ($G_{Prot} \cap G_{mRNA} = G_{Prot}$) common to all data sets. Thus, the differences between the populations now only reflect the effects of differential transcription of certain genes and differential translation of certain transcripts. We find here an enrichment specifically of cysteine in the translatome in relation to the transcriptome.

To measure the statistical significance of the results on amino acid enrichment, we have performed a control analysis on a randomized data set (Figure 2d). We randomly permutated the expression values of the ORFs 1000 times and then recomputed the enrichments. This allowed us to compute distributions for the amino acid enrichments and, from integrating these, one-sided p-values indicating the significance of the observed enrichments.

Amino acid enrichment in Transposon data set. We also tried to extend our methodology, ineffectively, to cope with the semi-quantitative Transposon set. We used only those 450 ORFs that consistently yielded either no expression or high expression, as binary data, on or off. We show the enrichments of amino acids computed from this filtered Transposon abundance set in Figure 2a. Overall, the enrichments from this set seemed to be attenuated in comparison to other data.

Biomass enrichment. A corollary to amino acid enrichments is the determination of the average biomass of the transcriptome and translatome populations (shown in Figure 2c). We found that the average molecular weight of a protein in both populations was, on average, lower than in the genome population. These preliminary observations suggest a cell preference to use less energetically expensive proteins for those that are highly transcribed or trans-

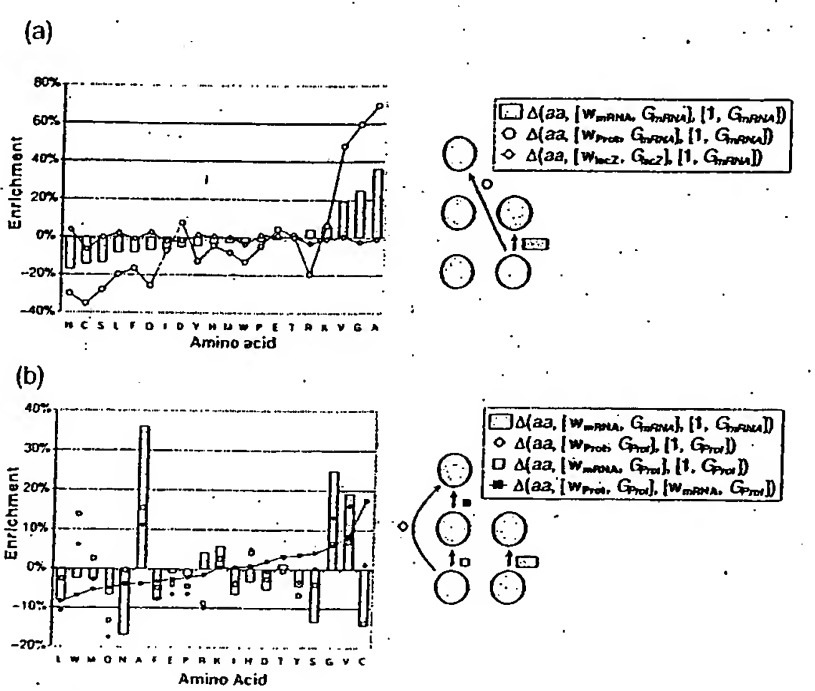


Fig. 2. Amino acid and biomass enrichment. (a) Shows the amino acid enrichments between different populations as indicated by the legend to the right of the plot (the legend is ordered in the same way as the schematic illustration in Figure 1). The bars indicate the enrichment of the transcriptome relative to the genome, whereas the circles indicate the enrichment of the translatome relative to the genome. In addition, we also show the enrichment for protein abundance from the Transposon abundance set, represented by the circles with the line through them. (b) Shows a different view of amino acid enrichment from that contained in (a), now focusing on changes, and thus restricting the comparison to the genes common to all the data sets. The graph is ordered according to the enrichment from transcriptome to translatome (black squares). We focus here only on the changes for the abundance gene set (G_{Prot}) to exclude the effects that arise from looking at different subsets. In this view the enrichments from genome to transcriptome (white squares) and from genome to translatome (white diamonds) look more similar than do the analogous sets in (a). To make comparison with (a) easier we again show the enrichment from genome to the transcriptome for the complete gene set (G_{Gen}, shown in bars). (c) Shows biomass enrichment. The left panel depicts the average molecular weight per ORF (in units of kDa) and the right panel, the average molecular weight per amino acid (in units of Daltons) in each of the three stages of gene expression. The numbers inside the circles indicate the average molecular weights. The values next to the arrows indicate the enrichments in biomass between different populations. Both the circle diameters and the arrow widths are functions of the corresponding values (the hollow arrow indicates a positive value). It is very clear that the average molecular weight per ORF is much lower in the translatome (by 20 or 15%) and transcriptome (by 29%) than in the genome. This relative depletion of biomass mainly takes place as a result of transcription; the effect of translation is less clear, depending on the populations compared. On the other hand, the depletion in the average molecular weight per amino acid (−3.3% from genome to translatome) is an order of magnitude smaller than in the average weight per ORF. This shows that the yeast cell favors the expression of shorter ORFs over longer ones, and agrees with our earlier observation that there is a negative correlation between maximum ORF length and mRNA expression (Jansen and Gerstein, 2000); it seems that this effect mainly takes place during transcription rather than translation. (d) This plot shows that the amino acid enrichments are statistically significant. We have assessed significance by randomly permuting the expression levels among the genes and then recomputing the amino acid enrichments. This procedure can be repeated and used to generate distributions of random enrichments that can then be compared against the observed enrichments. In the plot the gray bars represent the observed enrichments already shown in Figure 3a. On top of the gray bars we show standard boxplots of enrichment distributions based on 1000 random permutations. (The middle line represents the distribution median. The upper and lower sides of the box coincide with the upper and lower quartiles. Outliers are shown as dots and defined as data points that are outside the range of the whiskers, the length of which is 1.5 the interquartile distance.) Based on the random distributions, we can compute one-sided p-values for the observed enrichments. Amino acids for which the p-values are less than 10^{-3} are shown in bold font.

lated. However, we also found that the average molecular weight *per amino acid* differed much less between the transcriptome and the translatome on the one hand, and the genome on the other hand (though it was still slightly less). This finding indicates that lower molecular weights in the translatome and transcriptome relative to the genome are

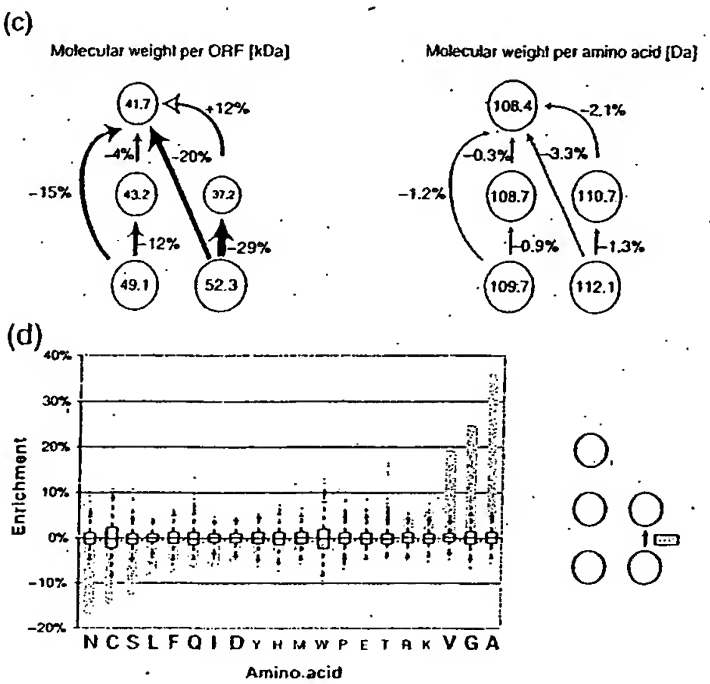


Fig. 2. Cont.

predominantly due to greater expression of shorter proteins rather than the incorporation of smaller amino acids.

Secondary structure composition. We also used our methodology to study the enrichment of secondary-structural features. Secondary structural annotation was derived from structure prediction applied uniformly to all the ORFs in the yeast genome as described in Table 1. As shown in Figure 3a, all three populations—genome, transcriptome, and translatome—had a fairly similar composition of secondary structures—sheets, helices, and coils. The differences between populations were marginal and based only on the small subset of genes.

We also found that Transmembrane (TM) proteins were significantly depleted in the transcriptome (see website and caption). These results are consistent with our previous analyses (Jansen and Gerstein, 2000). The protein abundance data does not have any membrane proteins.

Subcellular localization. Figure 3c shows the enrichment of proteins associated with the various subcellular compartments. For clarity, we divided the cell into five distinct subcellular compartments, (see Table 1). We found that, in comparison to the genome, both the transcriptome and translatome are enriched in cytoplasmic proteins. This is true whether we make our comparisons in relation to the relatively large reference mRNA expression set or the smaller reference protein abundance set. As Figure 3c shows, the 2D gel experiments are clearly biased towards proteins from the cytoplasm. However, in the biased subset G_{Prot} transcription and translation lead to an even higher fraction of cytoplasmic proteins in the translatome.

Functional categories. Finally, we compared the enrichment of various functional categories in both the translatome and the transcriptome (see Figure 3b). This gives us a broad yet informative view of the cell as a whole. As described in Table 1, we used the top-level of the MIPS scheme for the functional category definitions. We found broad differences between the various populations, with some of the functional categories showing strikingly high enrichments.

DISCUSSION AND CONCLUSION

We developed: (i) a methodology for integrating many different types of gene expression and protein abundance into a common framework and applied this to a preliminary analysis; (ii) a procedure for scaling and merging different mRNA and protein sets together; and (iii) an approach for computing the enrichment of various proteomic features in the population of transcripts and proteins. We showed that by analyzing broad categories instead of individual noisy

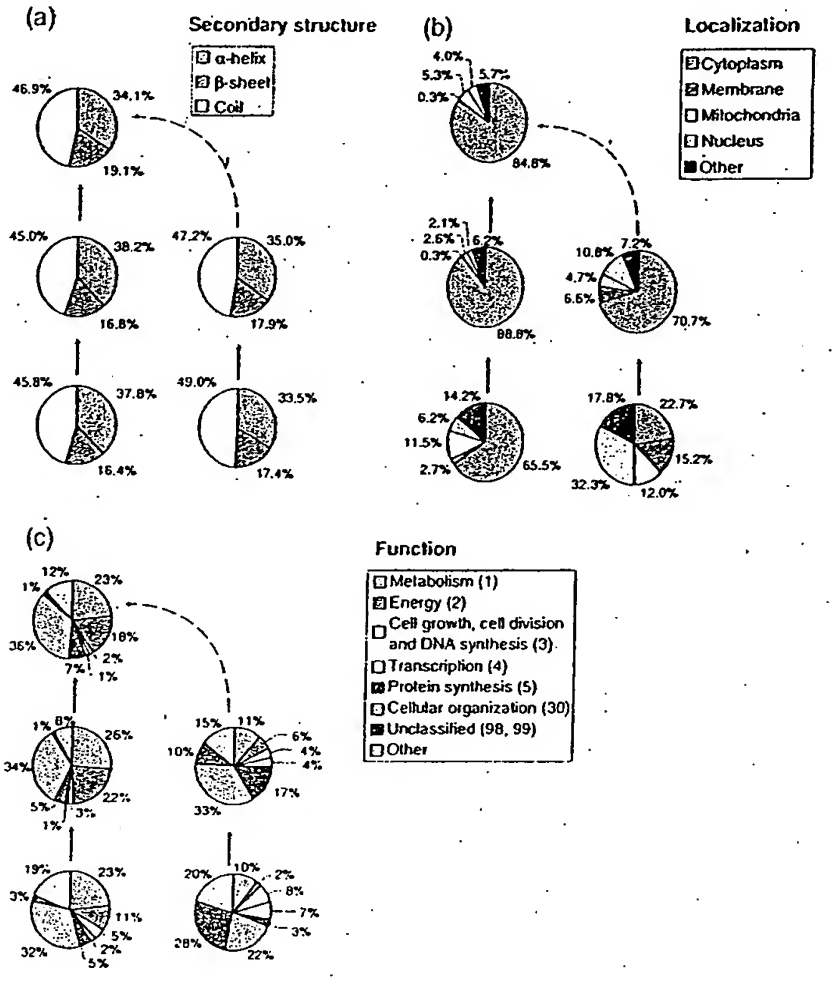


Fig. 3. Breakdown of the transcriptome and translatome in terms of broad categories relating to structure, localization, and function. All of the subfigures are analogous to the schematic illustration in Figure 1. (a) Represents the composition of secondary structure in the different populations. (b) Represents the distribution of subcellular localizations associated with proteins in the various populations. We used standardized localizations developed earlier (Drawid and Gerstein, 2000), which, in turn, were derived from the MIPS, YPD, and SwissProt databases (Bairoch and Apweiler, 2000; Costanzo *et al.*, 2000; Mewes *et al.*, 2000). The subcellular localization has been experimentally determined for less than half of the yeast proteins, so our analysis applies only to this subset. (c) Shows the division of ORFs into different functional categories (according to the MIPS classification) in the various populations. Only the largest functional categories of the top level of the MIPS classification are shown. The group 'other' contains the smaller top-level categories lumped together. This 'other' group is different from the group 'unclassified,' which contains genes without any functional description.

data points, we could find logical trends in the underlying data. For example, individual transcription factors might have higher or lower protein abundance than one expects from their mRNA expression, but the category 'transcription factors' as a whole has a similar representation in the transcriptome and translatome.

We found, as previously described (Futcher *et al.*, 1999; Gygi *et al.*, 1999b; Greenbaum *et al.*, 2001), a weak correlation between individual measurements of mRNA and protein abundance. The outliers of this correlation tend to be associated with cellular organization. One might conceive of using these outliers (i.e. those with significantly different transcriptional and translational behavior) to find consensus regulatory sequences. One possible method would involve using predicted mRNA structures (Jaeger *et al.*, 1990; Zuker, 2000) to find and investigate consensus structural elements in these outliers to which the yeast translational machinery is known to be

Table 1. Data sets

Data set	Description	Size [ORFs]	Reference
mRNA expression			
Young	Gene chip profiles yeast cells with mutations that affect transcription	5455	Holstege *et al.* (1998)
Church	Gene chip profiles of yeast cells under four different conditions	6263	Roth *et al.* (1998)
Samson	Comparing gene chip profiles for yeast cells subjected to alkylating agent	6090	Jelinsky and Samson (1999)
SAGE	Yeast cells during vegetative growth	3778	Velculescu *et al.* (1997)
Reference expression	Scaling and integrating the mRNA expression set into one data source	6249	–
Protein abundance			
2-DE #1	Measurement of yeast protein abundance by 2D gel electrophoresis and mass spectrometry	156	Gygi *et al.* (1999a,b)
2-DE #2	Similar to 2-DE set #1	71	Futcher *et al.* (1999)
Transposon	Large-scale fusions of yeast genes with *lacZ* by transposon insertion	1410	Ross-Macdonald *et al.* (1999)
Reference abundance	Scaling and integrating the 2-DE data sets into one data source	181	–
Annotation			
Annotated localization	Subcellular localizations of yeast proteins	2133 (6280)	Drawid and Gerstein (2000)
TM segments	Predicted TM and soluble proteins in yeast	2710 (6280)	Gerstein (1998a,b,c)
MIPS functions	Functional categories for yeast ORFs	3519 (6194)	Mewes *et al.* (2000)
GOR secondary structure	Predicted secondary structure yeast ORFs	6280	Gerstein (1998a,b,c)

This table provides an overview of the data sets used in our analysis. The table is divided into three sections. The top section lists different mRNA expression sets. The middle section shows the protein abundance data sets used. The bottom section contains different annotations of protein features. The column 'Data set' lists a shorthand reference to each data set used throughout this paper. The next columns contain a brief description of the data sets, the number of ORFs contained in each of them, and the literature reference. In contrast to the other data we investigated, the reference expression and abundance data sets have been calculated for the purpose of our analysis (see text). An expanded version of the table is available on our web site.

Some further information on the genome annotations:

Localization. Protein localization information from YPD, MIPS and SwissProt were merged, filtered and standardized (Bairoch and Apweiler, 2000; Costanzo *et al.*, 2000; Mewes *et al.*, 2000) into five simplified compartments—cytoplasm, nucleus, membrane, extracellular (including proteins in ER and golgi), and mitochondrial—according to the protocol in Drawid *et al.* (2000). This yielded a standardized annotation of protein subcellular localization for 2133 out of 6280 ORFs.

TM segments. In 2710 out of 6280 yeast ORFs TM segments are predicted to occur, ranging from low to high confidence (732 ORFs). The TM prediction was performed as follows: the values from the scale for amino acids in a window of size 20 (the typical size of a TM helix) were averaged and then compared against a cutoff of -1 kcal mol^{-1}. A value under this cutoff was taken to indicate the existence of a TM helix. Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first seven, followed by a stretch of 14 with an average hydrophobicity under the cutoff.) These parameters have been used, tested, and refined on surveys of membrane protein in genomes. 'Sure' membrane proteins had at least two TM-segments with an average hydrophobicity less than -2 kcal mol^{-1} (Rost *et al.*, 1995; Gerstein *et al.*, 2000; Santoni *et al.*, 2000; Senes *et al.*, 2000).

Functions. MIPS functional categories have been assigned to 3519 out of 6194 ORFs. (The remainder are assigned to category '98' or '99,' which corresponds to unclassified function.)

sensitive (McCarthy, 1998).

In relation to functional categories, we found three trends that were particularly notable: (i) the 'cellular organization,' 'protein synthesis,' and 'energy production' categories were increasingly enriched as we moved from genome to transcriptome to translatome. In the transcrip-

tome and translatome population relative to the genome; (ii) proteins with 'unclassified function' are significantly depleted, perhaps reflecting a bias against studying them; (iii) proteins in the 'transcription' and 'cell growth, cell division, and DNA synthesis' categories were consistently depleted. This reflects the fact that many of these proteins, such as transcription factors, act as 'switches' such that only small quantities of the protein are necessary to activate or deactivate a process. These results concur with previous calculations (Jansen and Gerstein, 2000) wherein we found the transcriptome is enriched specifically with proteins involved in protein synthesis and energy.

Limitations given the small size of the protein abundance data

Even with the extended coverage made possible by merging many data sets together into reference sets, the analysis is still limited by the minimal data. This was most applicable to the protein abundance measurements, potentially biasing our statistical results towards certain protein families. Moreover, the 181 proteins in G_{Prot} do not represent a random sample. They are skewed towards highly expressed, well-studied proteins. Our methodology attempts to control for this gene-selection bias through our enrichment formalism, which allows one to rather precisely gauge various aspects of the bias. Conversely, many protein features in both the translatome and the transcriptome are dominated by highly expressed proteins. Under these circumstances, it is often sufficient to look at this smaller number of dominating proteins to characterize the whole population. This is similar to the development of the codon adaptation index for yeast (Sharp and Li, 1987). While based on only 24 highly expressed proteins, it has proven to be robust in predicting expression levels for the entire genome.

We believe that the essential formalism and approach that we develop will remain quite relevant for future data sets (Smith, 2000).

ACKNOWLEDGEMENT

M.G. thanks the Keck foundation for support.

REFERENCES

An,H., Scopes,R.K. *et al.* (1991) Gel electrophoretic analysis of *Zymomonas mobilis* glycolytic and fermentative enzymes: identification of alcohol dehydrogenase II as a stress protein. *J. Bacteriol.*, 173, 5975–5982.

Anderson,L. and Seilhamer,J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, 18, 533–537.

Bairoch,A. (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics*, 16, 48–64.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28, 45–48.

Bassett,D.E. Jr., Basrai,M.A. *et al.* (1996) Exploiting the complete yeast genome sequence. *Curr. Opin. Genet. Dev.*, 6, 763–766.

Batke,J., Benito,V.A. *et al.* (1992) A possible *in vivo* mechanism of intermediate transfer by glycolytic enzyme complexes: steady state fluorescence anisotropy analysis of an enzyme complex formation. *Arch. Biochem. Biophys.*, 296, 654–659.

Cambillau,C. and Claverie,J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.*, 275, 32383–32386.

Cavalcoli,J.D., VanBogelen,R.A. *et al.* (1997) Unique identification of proteins from small genome organisms: theoretical feasibility of high throughput proteome analysis. *Electrophoresis*, 18, 2703–2708.

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression [in process citation]. *Hum. Mol. Genet.*, 8, 1821–1832.

Corthals,G., Wasinger,V.C., Hochstrasser,D.F. and Sanchez,J.C. (2000) The dynamic range of protein expression: a challenge for proteomic research. *Electrophoreisis*, 21, 1104–1115.

Costanzo,M.C., Hogan,J.D. *et al.* (2000) The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.*, 28, 73–76.

Das,R. and Gerstein,M. (2000) The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Funct. Int. Genom.*, 1, 33–45.

Doolittle,W.F. (2000) The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.*, 10, 355–358.

Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.*, 301, 1059–1075.

Drawid,A., Jansen,R. *et al.* (2000) Gene expression levels are correlated with protein subcellular localization. *Trends Genet.*, 10, 426–430.

Einarson,M. and Golemis,E. (2000) Encroaching genomics: adapting large-scale science to small academic laboratories. *Physiol. Genom.*, 2, 85–92.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Meth. Enzymol.*, 303, 179–205.

Epstein,C. and Butow,R. (2000) Microarray technology—enhanced versatility, persistent challenge. *Curr. Opin. Biotechnol.*, 11, 36–41.

Ferea,T. and Brown,P. (1999) Observing the living genome. *Curr. Opin. Genet. Dev.*, 9, 715–722.

Fey,S.J., Nawrocki,A. *et al.* (1997) Proteome analysis of *Saccharomyces cerevisiae*: a methodological outline. *Electrophoresis*, 18, 1361–72.

Fey,S.J. and Larsen,P.M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis. *Curr. Opin. Chem. Biol.*, 5, 26–33.

Frishman,D. and Mewes,H.W. (1997) Protein structural classes in five complete genomes [letter]. *Nat. Struct. Biol.*, 4, 626–628.

Frishman,D. and Mewes,H.W. (1999) Genome-based structural biology. *Prog. Biophys. Mol. Biol.*, 72, 1–17.

Futcher,B., Latter,G. *et al.* (1999) A sampling of the yeast proteome. *Mol. Cell Biol.*, **19**, 7357–7368.

Gaasterland,T. (1999) Archaeal genomics. *Curr. Opin. Microbiol.*, **2**, 542–547.

Garrels,J.I., McLaughlin,C.S. *et al.* (1997) Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis*, **18**, 1347–1360.

Gerstein,M. (1997) A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.*, **274**, 562–576.

Gerstein,M. (1998a) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.*, **3**, 497–512.

Gerstein,M. (1998b) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, **33**, 518–534.

Gerstein,M. (1998c) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, **33**, 518–534.

Gerstein,M. and Hegyi,H. (1998) Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.*, **22**, 277–304.

Gerstein,M. and Jansen,R. (2000) The current excitement in bioinformatics, analysis of whole-genome expression data: how does it relate to protein structure and function. *Curr. Opin. Struct. Biol.*, **10**, 574–584.

Gerstein,M., Lin,J. *et al.* (2000) Protein folds in the worm genome. *Pac. Symp. Biocomput.*, 30–41.

Greenbaum,D., Luscombe,N. *et al.* (2001) Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.*, **11**, 1463–1468.

Gygi,S.P., Rist,B. *et al.* (1999a) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.*, **17**, 994–999.

Gygi,S.P., Rochon,Y. *et al.* (1999b) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.*, **19**, 1720–1730.

Gygi,S.P., Corthals,G.L. *et al.* (2000a) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci. USA*, **97**, 9390–9395.

Gygi,S.P., Rist,B. *et al.* (2000b) Measuring gene expression by quantitative proteome analysis. *Curr. Opin. Biotechnol.*, **11**, 396–401.

Harry,J.L., Wilkins,M.R. *et al.* (2000) Proteomics: capacity versus utility. *Electrophoresis*, **21**, 1071–1081.

Hatzimanikatis,V., Choe,L.H. *et al.* (1999) Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.*, **15**, 312–318.

Haynes,P.A. and Yates,J.R. (2000) Proteome profiling-pitfalls and progress. *Yeast*, **17**, 81–87.

Hegyi,H. and Gerstein,M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.*, **288**, 147–164.

Holstege,F.C., Jennings,E.G. *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

Ishii,M., Hashimoto,S. *et al.* (2000) Direct comparison of genechip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, **68**, 136–143.

Ito,T., Tashiro,K. *et al.* (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.

Jaeger,J.A., Turner,D.H. *et al.* (1990) Predicting optimal and suboptimal secondary structure for RNA. *Meth. Enzymol.*, **183**, 281–306.

Jansen,R. and Gerstein,M. (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res.*, **28**, 1481–1488.

Jelinsky,S.A. and Samson,L.D. (1999) Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl Acad. Sci. USA*, **96**, 1486–1491.

Jones,D.T. (1998) Do transmembrane protein superfolds exist? *FEBS Lett.*, **423**, 281–285.

Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

Kidd,D. *et al.* (2001) Profiling serine hydrolase activities in complex proteomes. *Biochemistry*, **40**, 4005–4015.

Klose,J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, **26**, 231–243.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.*, **10**, 808–818.

Lipshutz,R.F. S., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet.*, **21**, 20–24.

Lopez,M.F. (2000) Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoreisis*, **21**, 1082–1093.

MacBeath,G. and Schreiber,S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

Matton,D.P., Constabel,P. *et al.* (1990) Alcohol dehydrogenase gene expression in potato following elicitor and stress treatment. *Plant Mol. Biol.*, **14**, 775–783.

McCarthy,J.E. (1998) Posttranscriptional control of gene expression in yeast. *Microbiol. Mol. Biol. Rev.*, **62**, 1492–1553.

Mewes,H.W., Frishman,D. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 27–40.

Millar,A.A., Olive,M.R. *et al.* (1994) The expression and anaerobic induction of alcohol dehydrogenase in cotton. *Biochem. Genet.*, **32**, 279–300.

Molloy,M.P. (2000) Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.*, **280**, 1–10.

Nauchitel,V.V. and Somorjai,R.L. (1994). Spatial and free energy distribution patterns of amino acid residues in water soluble proteins. *Biophys. Chem.*, **51**, 327–336.

Nelson,R.W., Nedelkov,D. *et al.* (2000) Biosensor chip mass spectrometry: a chip-based proteomics approach. *Electrophoresis*, **21**, 1155–1163.

O'Farrell,P.H. (1975) High resolution two-dimensional elec-

trophoresis of proteins. *J. Biol. Chem.*, 250, 4007–4021.
Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, 405, 837–846.
Qi,S.Y., Moir,A. *et al.* (1996) Proteome of *Salmonella typhimurium* SL1344: identification of novel abundant cell envelope proteins and assignment to a two-dimensional reference map. *J. Bacteriol.*, 178, 5032–5038.
Ross-Macdonald,P., Coelho,P.S. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402, 413–418.
Rost,B., Casadio,R. *et al.* (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci.*, 4, 521–533.
Roth,F.P., Hughes,J.D. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, 16, 939–945.
Rubin,G.M., Yandell,M.D. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, 287, 2204–2215.
Sali,A. (1999) Functional links between proteins. *Nature*, 402, 25–26.
Santoni,V., Molloy,M. *et al.* (2000) Membrane proteins and proteomics: un amour impossible? *Electrophoreisis*, 21, 1054–1070.
Schena,M., Shalon,D. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467–470.
Searls,D.B. (2000) Using bioinformatics in gene and drug discovery. *Drug Discov. Today*, 5, 135–143.
Senes,A., Gerstein,M. *et al.* (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.*, 296, 921–936.
Shapiro,L. and Harris,T. (2000) Finding function through structural genomics. *Curr. Opin. Biotechnol.*, 11, 31–35.
Sharp,P.M. and Li,W.H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, 15, 1281–1295.
Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, 12, 201–205.
Shevchenko,A., Jensen,O.N. *et al.* (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA*, 93, 14 440–14 445.
Smith,R.D. (2000) Probing proteomes-seeing the whole picture? *Nature Biotechnol.*, 18, 1041–1042.
Tatusov,R.L., Koonin,E.V. *et al.* (1997) A genomic perspective on protein families. *Science*, 278, 631–637.
Tekaia,F., Lazcano,A. *et al.* (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, 9, 550–557.
Velculescu,V.E., Zhang,L. *et al.* (1997) Characterization of the yeast transcriptome. *Cell*, 88, 243–251.
Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, 7, 1029–1038.
Washburn,M.P., Wolters,D. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.*, 19, 242–247.
Washburn,M.P. and Yates,J.R. 3rd (2000) Analysis of the microbial proteome. *Curr. Opin. Microbiol.*, 3, 292–297.
Wittes,J. and Friedman,H.P. (1999) Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data [editorial; comment]. *J. Natl Cancer. Inst.*, 91, 400–401.
Wolf,Y.I., Brenner,S.E. *et al.* (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res.*, 9, 17–26.
Young,K.H. (1998) Yeast two-hybrid: so many interactions, (in) so little time ...: *Biol. Reprod.*, 58, 302–311.
Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologists (published erratum appears in *Genome Res.*, 1999, 9, 1156). *Genome Res.*, 9, 681–688.
Zhu,H., Klemic,J.F. *et al.* (2000) Analysis of yeast protein kinases using protein chips. *Nature Genet.*, 26, 283–289.
Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, 10, 303–310.

Review

Early Detection of Lung Cancer: Clinical Perspectives of Recent Advances in Biology and Radiology[1]

Fred R. Hirsch,[2] Wilbur A. Franklin, Adi F. Gazdar, and Paul A. Bunn, Jr.

Lung Cancer Program and Departments of Medicine and Pathology, University of Colorado Cancer Center, Denver, Colorado 80262 [F. R. H., W. A. F., P. A. B.]; Department of Pathology, University of Texas, Southwestern Medical Center, Dallas, Texas [A. F. G.]; and Department of Oncology, Finsen Center, National University Hospital, Copenhagen, Denmark [F. R. H.]

Abstract

Lung cancer is the most common cause of cancer death in developed countries. The prognosis is poor, with less than 15% of patients surviving 5 years after diagnosis. The poor prognosis is attributable to lack of efficient diagnostic methods for early detection and lack of successful treatment for metastatic disease. Most patients (>75%) present with stage III or IV disease and are rarely curable with current therapies. Within the last decade, rapid advances in molecular biology, pathology, bronchology, and radiology have provided a rational basis for improving outcome. These advancements have led to a better documentation of morphological changes in the bronchial epithelium before development of clinical evident invasive carcinomas. This has changed our concept of lung carcinogenesis and emphasized the multistep carcinogenesis approach on several levels. Combined with the technical developments in bronchoscopic techniques, *e.g.*, laser-induced fluorescence endoscope (LIFE) bronchoscopy, we now have improved methods to localize preinvasive and early-invasive bronchial lesions. With the LIFE bronchoscope, a new morphological entity (angiogenic squamous dysplasia) has been recognized, which might be an important biomarker and target for antiangiogenic chemopreventive agents. To reduce the mortality of lung cancer, these new technologies have been taken into the clinic in different scientific settings. The use of low-dose spiral computed tomography in the screening of a high-risk population has demonstrated the possibility of diagnosing small peripheral tumors that are not seen on conventional X-ray. A shift in the therapeutic paradigm from targeting advanced clinically manifest lung cancer toward asymptomatic preinvasive and early-invasive cancer is occurring. The present article reviews the recent advances in the diagnosis of preinvasive and early-invasive cancer to identify biomarkers for early detection of lung cancer and for chemoprevention studies.

Introduction

Lung cancer is the most common cause of cancer deaths in the countries of North America and other developed countries, accounting for 29% of all cancer deaths and more deaths than from prostate, breast, and colorectal cancer combined in the United States (1). Lung cancer will be diagnosed in ~170,000 new patients in the United States in the year 2000, and <15% of them will survive 5 years after diagnosis (1). The prognosis for the patients with lung cancer is strongly correlated to the stage of the disease at the time of diagnosis. Whereas patients with clinical stage IA disease have a 5-year survival of about 60%, the clinical stage II-IV disease 5-year survival rate ranges from 40% to less than 5% (2). Over two-thirds of the patients have regional lymph-node involvement or distant disease at the time of presentation (3). The poor prognosis is largely attributable to the lack of effective early detection methods and the inability to cure metastatic disease. The unsatisfactory cure rates supports efforts aimed at early identification and intervention in lung cancer.

Historically, the only diagnostic tests available for the detection of lung cancer in its early stages were chest radiography and sputum cytology. The efficacy of these tests as mass screening tools was evaluated in controlled trials sponsored by the NCI[3] and conducted at Johns Hopkins University, Memorial Sloan-Kettering Cancer Center, and the Mayo Clinic during the 1970s (4–6). The principal goal of these studies was to determine whether a reduction in lung cancer mortality could be achieved by adding sputum cytology testing to annual screening by chest radiography. Results from these trials showed that both tests could detect presymptomatic, early-stage carcinoma, particularly of squamous cell type. Resectability and survival rates were found to be generally higher in the study groups than in the control groups. However, improvements in resectability and survival did not lead to a reduction in overall lung cancer mortality, the most critical end point. A subsequent study of 6346 Czechoslovakian male smokers also found no reduction in lung cancer mortality after dual screening by chest radiography

Received 6/29/00; revised 10/16/00; accepted 10/30/00.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] Supported by National Cancer Institute Grants CA 58187 from the Specialized Program of Research Excellence (SPORE)-Lung and CA 85070 from the Lung Cancer Biomarkers and Chemoprevention Consortium.

[2] To whom requests for reprints should be addressed, at University of Colorado Cancer Center, Department of Pathology, University of Colorado Health Sciences Center, 4200 East Ninth Avenue, B 216, Denver, Colorado 80262. E-mail: Fred.Hirsch@UCHSC.edu

[3] The abbreviations used are: NCI, National Cancer Institute; CIS, carcinoma *in situ*; CT, computed tomography; ASD, angiogenic squamous dysplasia; TSG, tumor suppressor gene; LOH, loss of heterozygosity; hnRNP, heterogeneous nuclear ribonucleoprotein; SPLC, second primary lung cancer; BAL, bronchoalveolar lavage; SCLC, small cell lung carcinoma; WLB, white light bronchoscopy; LIFE, laser-induced fluorescence endoscope; ELCAP, Early Lung Cancer Action Project; PET, positron emission tomography; FDG, [^{18}F]fluoro-2-deoxyglucose.

and sputum cytology (7). The negative results from these screening studies lead the NCI and other health policy and research groups to conclude that mass screening programs involving periodic sputum cytological evaluation and chest radiographs could not be justified. However, controversies in the methodology and interpretation of the data from these studies have later been extensively discussed (8, 9). One additional study of annual chest X-ray screening is currently being conducted by the NCI; The Prostate-, Lung-, Colorectal-, and Ovarian (PLCO) screening trial. This trial includes individuals 55–74 years old, but they are not selected for this trial on the basis of high risk for lung cancer (*e.g.*, smoking history with >20 pack-years).

The failure of clinical trials to demonstrate the efficacy of sputum cytology and chest radiography as mass screening tools has resulted in a search for better diagnostic approaches for early lung cancer detection that take advantage of recent developments in molecular biology, gene technology, and radiology (10). Furthermore, as has been the case for mammography screening for breast cancer, it has also been important to identify risk groups for lung cancer.

Although, much is known about predisposing factors, natural history, and the outcome based on histology and stage, our understanding remains very incomplete in many areas. What are the early premalignant changes molecularly, biochemically, and morphologically? Which changes are reversible and which are not? What research tools are available to provide answers to these questions? The identification of preinvasive lesions allows for developing promising methods for early intervention (11). The therapeutic paradigm and focus are today shifting from targeting only clinically verified lung cancer as previously toward targeting the premalignant and early- malignant lesions. Furthermore, the prospect of lung cancer screening has today become more meaningful as a consequence of recent developments in biology and radiology and better possibilities to define high-risk populations most suitable for lung cancer screening (12).

The present article will focus on the clinical perspectives of our biological knowledge of premalignant and early-malignant lesions and the potential of the recent technological advancement for early diagnosis of lung cancer.

Pathology of Preinvasive and Early Invasive Bronchial Lesions

Most of the efforts to classify lung cancer have been directed toward invasive carcinoma (13). However, better understanding of the pathogenesis of lung cancer aroused renewed interest in morphological abnormalities that fall short of invasive carcinoma but may indicate initiation of carcinogenesis. These morphological abnormalities are referred to as preinvasive lesions and are shown in Fig. 1. The last edition of the WHO classification of lung tumors included the classification of preinvasive lesions as a separate section. Numerous recent studies have indicated that lung cancer is not the result of a sudden transforming event in the bronchial epithelium but a multistep process in which gradually accruing sequential genetic and cellular changes result in the formation of an invasive (*i.e.*, malignant) tumor. Mucosal changes in the large airways, that may precede or accompany invasive squamous carcinoma include hyperplasia, metaplasia, dysplasia, and CIS (14). Hyperplasia of the bronchial epithelium and squamous metaplasia have generally been considered reversible, and not premalignant in the sense of squamous dysplasia and CIS (15).

Squamous metaplasia is a common finding, especially as a response to cigarette smoking. Peters *et al.* (16) studied bronchoscopic biopsies from six sites in 106 heavy cigarette smokers; Squamous metaplasia was noted at one or more biopsy sites in approximately two- thirds of the group, and one-fourth showed squamous metaplasia in three or more biopsy sites. The incidence of squamous metaplasia increased with smoking history and was highest in individuals who had smoked more than two packs of cigarettes a day. Auerbach *et al.* (17) noted similar findings in autopsy tissues: basal cell hyperplasia and squamous metaplasia are increased in smokers in proportion to smoking history. Hyperplasia and metaplasia are believed to be reactive changes in the bronchial epithelium, as opposed to true preneoplastic changes (17, 18). The reasons for this include: (*a*) they are frequently found in association with chronic inflammation, and may be induced by mechanical trauma; (*b*) they spontaneously regress after smoking cessation; (*c*) in chronic smokers, the molecular changes present in these lesions are similar to those present in histologically normal epithelium; and (*d*) there are no reports linking their presence to increased risk for developing lung cancer. In contrast, moderate-to-severe dysplasia and CIS lesions seldom regress after smoking cessation (19).

Dysplasia and CIS are changes that frequently precede squamous cell carcinoma of the lung. Saccomanno *et al.* (20) studied more than 50,000 samples from 6,000 men, many of whom had worked in the uranium mining industry. Both smoking and uranium mining (radon exposure) were found to be associated with increased incidence of dysplasia, CIS, and invasive cancer. The studies of Saccomanno *et al.* established that increasing degrees of sputum atypia may be recognized an average of 4–5 years before the development of frank lung carcinoma.

Another question is: which grades of sputum atypia progress to cancer? From the Johns Hopkins cohort of the NCI chest X-ray/sputum screening trial, we know that among individuals with moderate atypia on sputum screening, ~10% developed known cancer up to 9 years later. Among individuals with severe atypia on the sputum screening, >40% developed known cancer during the same time period (21). Although there are data in the literature showing the relationship between sputum atypia and subsequent invasive cancer, there is still very little information about the histological progression in the bronchial mucosa in the high risk populations. In a recent publication, nine patients with CIS were followed with autofluorescence bronchoscopy at regular intervals, and 5 (56%) had progression to invasive cancer despite endobronchial therapy (22). The number of invasive cancers might even have been higher if treatment had not been not given. Ongoing studies of high-risk subjects (*e.g.*, the Colorado sputum cohort study) including serial follow-up bronchoscopies will provide evidence related to the frequency of development of invasive lung cancer as it relates to smoking history, airflow obstruction, and sputum atypia.

Since the previous WHO-classification was published in

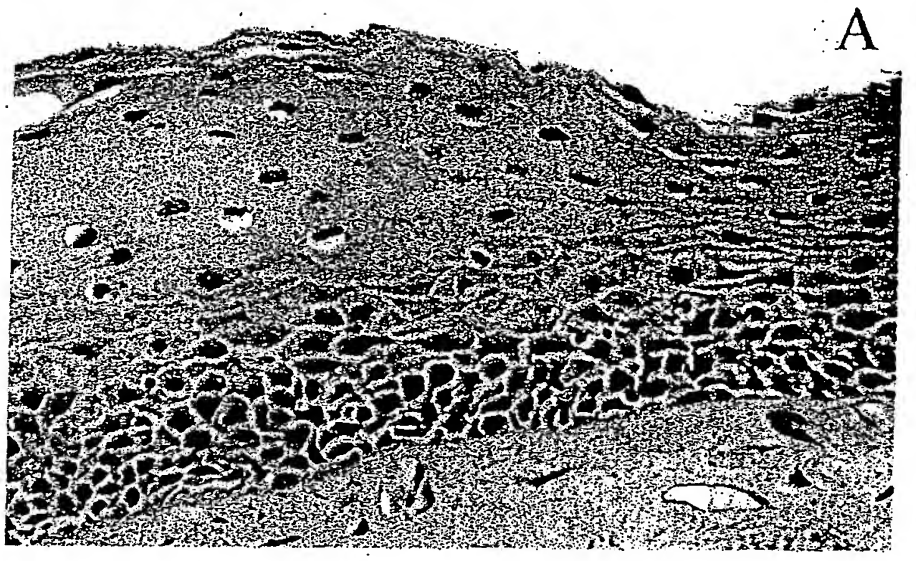


Fig. 1 *A*, squamous metaplasia. The cells are widely dispersed, with a regular maturation from the basal region to the top. There is keratinization, and the nuclei/cytoplasmic ratio is low. *B*, moderate dysplasia with ASD. Hypercellularity of the epithelium with incomplete maturation and micropapillary invasion of capillaries are seen. The nuclei/cytoplasmic ratio is high. *C*, severe dysplasia. There is marked pleomorphism of the cells with irregularity and prominent nucleoli.

1981, two nonsquamous lesions have been added to the WHO classification of premalignant lesions: atypical alveolar hyperplasia and diffuse idiopathic neuroendocrine cell hyperplasia (13). Both of these lesions are diagnosed rarely. The former consists of lesions <5 mm in diameter and composed of a peripheral epithelial cell proliferation with minimal cytological atypia or stromal response and resembles bronchioloalveolar carcinoma. The lesion has been seen in lung specimens resected for lung cancer, but no prospective significance of this lesion has been reported. However, this morphological lesion may play a role for the pathogenesis of peripheral lung adenocarcinomas (23, 24). The resolution of spiral CT (currently about 3 mm) approaches the diameter of these lesions, and it is anticipated that atypical alveolar hyperplasia will be increasingly encountered in subjects undergoing this procedure (25). Diffuse idiopathic neuroendocrine cell hyperplasia consists of a patchy increase in the number of well-differentiated neuroendocrine cells in the bronchioles. This process may result in the formation of small carcinoid tumors, and for this reason it is considered "preinvasive." To date, small cell carcinomas have not been associated with this lesion (13).

Recently, the use of fluorescence bronchoscopy (see below) has increased the recognition of dysplastic lesions in the large airways and a new morphological entity, ASD, was identified (26). Dysplasia of bronchial epithelium in "micropapillomatosis" and the possible link between angiogenesis and preinvasive bronchial epithelial dysplasia were recognized as early as 1983 by Muller and Muller (27), who also described the ultrastructure of these lesions. It has been suggested that this angiogenesis, which is recognized as capillary loops projecting into the dysplastic bronchial lining, is responsible for the reduced fluorescence seen in dysplastic lesions by LIFE bronchoscopes (Figs. 1 and 3; Ref. 26). Future prospective studies will show whether this morphological entity is correlated with a progression to lung cancer so as to be a target for the use of antiangiogenic agents for chemoprevention.

In general, there are several questions/problems relating to premalignant lesions, which will be addressed in future studies:

(*a*) The morphological criteria for premalignant and early-malignant changes, both on sputum cytology and in bronchial biopsies, have to be validated for intra- and interobserver reproducibility.

(*b*) Uniform and reproducible morphological/cytological criteria have to be published more extensively, and a training set of slides should be available. By the use of Internet technology, this could be more easily facilitated (28).

(*c*) The correlation of sputum atypia and histological changes in the bronchi in high-risk population is not well defined.

(*d*) The natural course of preinvasive changes in the bronchi from the high risk subjects needs to be clarified through longitudinal, prospective studies with reference to histological changes in the bronchi. Ongoing longitudinal studies with fluorescence bronchoscopy and multiple biopsies with histology and other biomarkers will define the ability of these markers to assess for risk.

(*e*) What is the pathology-biology of the small, often peripherally located, tumors (3 mm in diameter), which are more often diagnosed with newer radiological techniques (*e.g.*, low-dose spiral CT)?

(*f*) Optimization of the tissue procurement and processing techniques are important. Distinction of reactive from neoplastic processes is usually straightforward, but diagnostic difficulties may arise in the case of (*a*) inadequate or poorly prepared histological material to evaluate and (*b*) the presence of cytological atypia in epithelium stimulated by inflammation, viral infection, radiation, or chemotherapy.

(*g*) DNA array analyses of gene expression: will it be useful? How to collect proper mRNA? Can mRNA extracted from microdissected cells obtained at bronchoscopy be globally amplified and still remain representative of RNA present *in situ*?

Biology of Lung Carcinogenesis and Potential Early Detection Markers

Lung cancer is the end-stage of multiple-step carcinogenesis, in most cases driven by genetic and epigenetic damage caused by chronic exposure to tobacco carcinogens. The genetic instability in human cancers appears to exist at two levels: at the chromosomal level, including large scale losses and gains; and at the nucleotide level including single or several base changes (29). Lung cancers harbor many numerical chromosomal abnormalities (aneuploidy) and structural cytogenetic abnormalities including deletions and nonreciprocal translocations (30). At least three classes of cellular genes are involved: proto-oncogenes, TSGs, and DNA repair genes. Oncogenic activation often occurs via point mutations, gene amplification, or chromosomal rearrangement, whereas TSGs are classically inactivated by the loss of one parental allele combined with a point or small mutation or aberrant methylation of a target TSG in the remaining allele. Additionally, dysregulated gene expression (either increased or decreased expression) can occur by other, as yet unknown, mechanisms (30). Present studies have not yet confirmed a prominent role for abnormalities of DNA repair genes in lung cancer.

Preneoplastic cells contain several molecular genetic abnormalities identical to some of the abnormalities found in overt lung cancer cells (Fig. 2). These include allele loss at several loci (*3p*, *9p*, *8p*, and *17p*), *myc* and *ras* up-regulation, cyclin D1 overexpression, p53 mutations, and increased immunoreactivity, bcl-2 overexpression and DNA aneuploidy (31–35). Allelotyping of precisely microdissected, preneoplastic foci of cells suggests that the earliest changes in the bronchial epithelium is allele loss at chromosome regions 3p, then 9p, 8p, 17p, 5q, and then *ras* mutations (36–39).The biological meaning of LOH is only vaguely understood. Recent evidence suggests that LOH may be a consequence of mitotic recombination, that there is only infrequent physical loss of genetic loci, and that LOH probably precedes chromosomal duplication (40). Allelic loss would thus be significant primarily in the presence of mutation in the retained allele, and gene dosage would not be expected to exert a phenotypic effect in LOH. Some reports have indicated that *ras* activation occurs at early carcinoma stages (34). Histologically normal bronchial epithelium adjacent to cancers has also been shown to have certain genetic losses. Atypical adenomatous hyperplasia, the potential precursor lesion of adenocarcinomas, often have Ki-*ras* mutations (41).

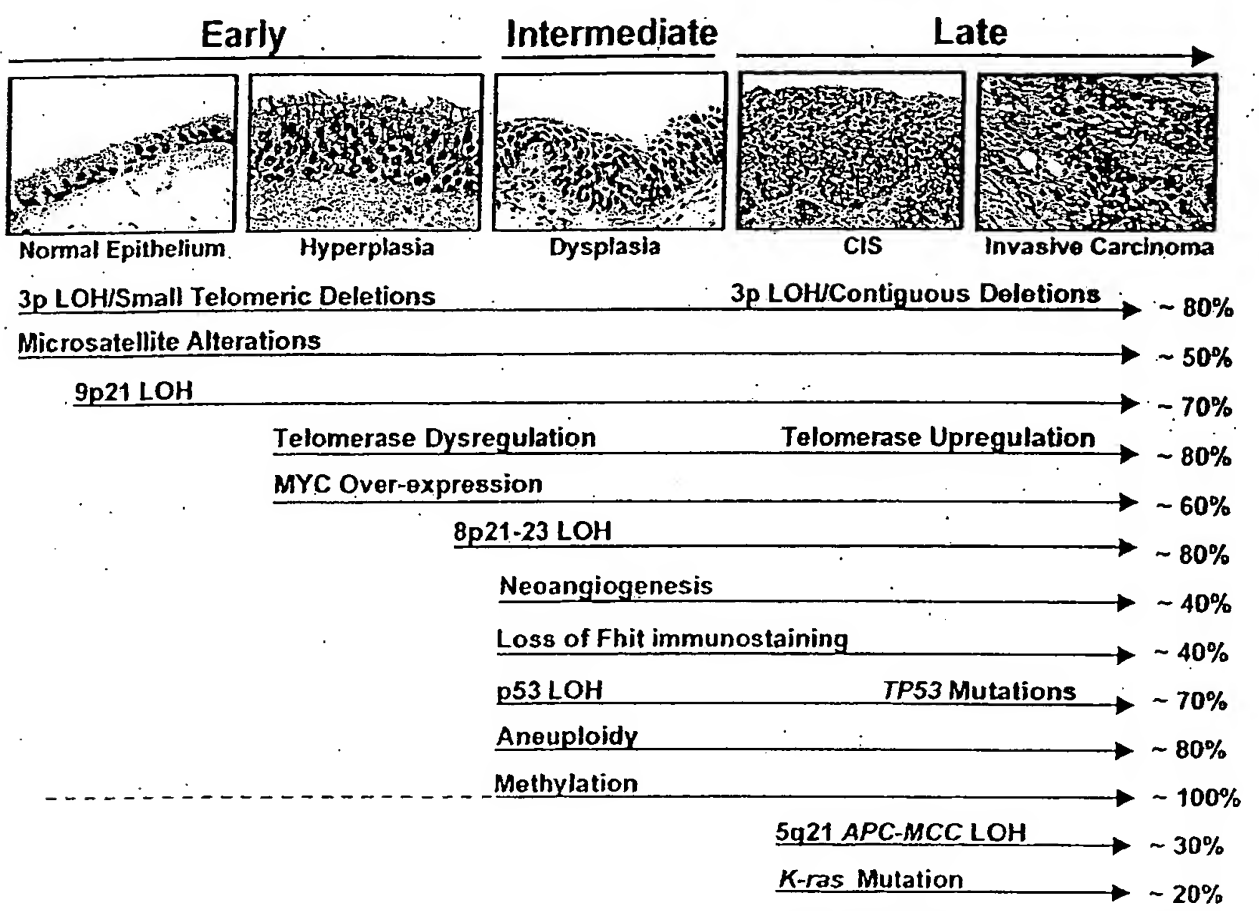


Fig. 2 Sequential changes during lung cancer pathogenesis. Although multiple genetic markers are abnormal in lung cancers, their appearance during the lengthy preneoplastic process varies. The timing of the appearance of these changes has been investigated in bronchial preneoplasia, because sequential sampling of the peripheral lung is technically difficult. Several alterations have been described in histologically normal bronchial epithelium of smokers. Other changes have been detected in slightly abnormal epithelium (hyperplasia, metaplasia) which we do not consider to be true premalignant lesions. These changes are regarded as early changes. Molecular changes detected frequently in dysplasia are regarded as intermediate in timing, whereas those usually detected at the CIS or invasive stages are regarded as late changes. It should be stressed that although there is a usual order, exceptions regarding the timing of onset may occur. Some changes are progressive, such as chromosome 3p deletions. Thus small discrete changes are present early, progressively become more extensive during pathogenesis, and frequently involve all or almost all of the arm in CIS samples. Although allelic loss at the TP53 locus may precede the onset of mutations, data on this sequence are scant. Dysregulation of the RNA component of telomerase (with its appearance in nonbasal cells) is an early event, whereas up-regulation of the gene is a relatively late event.

Molecular changes have been found not only in the lungs of patients with lung cancer, but also in the lungs of current and former smokers without lung cancer (18, 42, 43). These observations are consistent with the multistep model of carcinogenesis and "field cancerization" process, whereby the whole region is repeatedly exposed to carcinogenic damage (tobacco smoke) and is at risk for developing multiple, separate, clonally unrelated foci of neoplasia. The widespread aneuploidy that occurs throughout the respiratory tree of smokers supports this theory (44). However, the presence of the same somatic p53 point mutation at widely dispersed preneoplastic lesions in a smoker without invasive lung cancer indicates that expansion of a single progenitor clone may spread throughout the respiratory tree (45). These molecular alterations might thus be important targets for use in the early detection of lung cancer and for use as surrogate biomarkers in the follow-up of chemoprevention studies. Detection of these mutant cells should be possible with the different molecular techniques in accessible specimens. The prospects of diagnosing these biological abnormalities in multiple types of clinical specimens are discussed below.

Specimens for Clinical Testing: Sputum

Since the 1930s, cytological examination of sputum has been used for the diagnosis of lung cancer (46). Cytological examination of sputa, especially multiple samples, is helpful for the detection of central tumors arising from the larger bronchi (*e.g.*, squamous cell- and small cell carcinomas). Exfoliated cells from peripheral tumors, such as adenocarcinomas, arising from the smaller airways (small bronchi, bronchioles, and alveoli), especially those less than 2 cm in diameter, can be detected only occasionally in sputum samples. This has become of greater importance because the changes in cigarette exposition

(filters and decreased nicotine content) have created an increase in adenocarcinomas and a decrease in squamous carcinomas (47–49). The sensitivity of sputum cytology for early lung cancer is only in the 20%–30% range from screening studies, but by adhering to proper specimen collection, and processing and interpreting criteria, the yield can be substantially improved (50, 51). The data on the reliability of the sputum are conflicting (52–54). Browman *et al.* (52) reported interobserver agreement of 68% for exact and 82% for within - 1–category. Holliday *et al.* (54) reported low agreement within observers (27–60%) and across observers (13–50%). Within - 1 - category intraobserver agreement underwent a two- or 3-fold increase in agreement, which was also the case for interobserver agreement. The variation in intra- and interobserver agreement seems to depend on experience among the cytotechnicians/cytopathologists and the composition of categories studied. A higher degree of agreement is obtained for higher grades of dysplasia (54). Risse *et al.* (55) showed that the ability to detect premalignant conditions is dependent on the number and type of cells present in the deeper airways, suggesting a mode of improvement that is unrelated to observer reliability. MacDougall *et al.* (56) concluded that sputum cytology was too insensitive and insufficiently accurate to be included in the routine work-up of any patient suspected of having lung cancer. To improve the reliability of sputum cytology examinations a simplification of the diagnostic categories from 6 (normal; squamous metaplasia; mild, moderate, and severe atypia; and carcinoma) to 2–3 categories have been proposed (54). Future clinicopathological studies will be required to validate this concept.

To improve the sensitivity of sputum examination as a population-screening tool for the detection of early lung cancer, several approaches are currently under development.

Immunostaining. Annual sputum specimens obtained from individuals screened at Johns Hopkins were obtained, and the patients were monitored for 8 years (57). Because the clinical outcome of these patients was known, archival sputum specimens were screened for the presence of biomarkers that could indicate the presence of lung tumors in an early, preinvasive stage. In an attempt to distinguish the pattern of marker expression Tockman *et al.* (58) studied two monoclonal antibodies. Positive staining predicted subsequent lung cancer approximately 2 years before clinical recognition of the disease, with a sensitivity of 91% and a specificity of 88% (58). One of these antibodies (703 D4) had a higher sensitivity and was later identified as recognizing hnRNP A2/B1 (59). The role of hnRNP A2/B1 overexpression for detecting preclinical lung cancer has been studied in a large high-risk population including 6000 Chinese tin miners who were heavy smokers and who had an extraordinary rate of lung cancer (60). The results from this study indicated that detection of hnRNP A2/B1 overexpression in sputum epithelium cells was 2- to 3-fold more sensitive for detection of lung cancer than standard chest X-ray and sputum cytology methods. The method was particularly effective in identifying early disease (60). The sensitivity was 74% *versus* 21% for cytology and 42% for chest X-ray. However, the biomarker had a lower specificity (70%) compared with cytology (100%) and chest radiograph (90%). An ongoing clinical trial is evaluating the performance of the A2/B1 protein as a biomarker for the early detection of SPLC. The patients at risk for SPLC have the highest incidence of lung cancer (2–5%) among asymptomatic populations (61). In this trial, 13 SPLCs were identified by A2/B1, and the sensitivity and specificity were 77–82% and 65–81%, respectively. Among the cases identified as positive by immunocytochemistry and image cytometry, 67% developed SPLC within 1 year (62). Whereas the previous immunocytochemistry studies on material from the older screening material from the NCI-supported screening studies were made on sputum cells cytologically classified with moderately or gravely atypical metaplastic appearance, the latter studies have been done on cytologically "normal appearing" cells. More recently Sueoka *et al.* (63) reported the confirmation of the value of overexpression of hnRNP A2/B1 to detect preclinical lung cancer in Japan. Efforts to improve the sensitivity of hnRNP markers are ongoing (64).

PCR Techniques. PCR techniques have been used for the evaluation of molecular biomarkers for early lung cancer detection. In a pilot study with selected patients from the Johns Hopkins Lung Project (JHLP), 8 (53%) of 15 patients with adenocarcinoma or large cell carcinoma were detected by mutations in sputum cells from 1 to 13 months before clinical diagnosis (65). However, the method seemed to be less sensitive than the protein marker described above, and the identification of specific gene abnormalities is further limited by the need to know the specific mutation sequence with which to probe the sputum specimens. Currently, this approach is not practical for screening undiagnosed individuals. Future advances in gene chip technology may permit testing for all possible mutations of common oncogenes and TSGs in clinical specimens of asymptomatic individuals (62).

Microsatellite markers are small repeating DNA sequences found in the noncoding regions of a gene. PCR amplification of these repeat sequences provides a rapid method for assessment of LOH and facilitates the mapping of suppressor genes (66, 67). Microsatellite alterations are extension or deletions of these repeated elements. Detection of microsatellite alterations in histological or cytological specimens may facilitate the detection of clonal preneoplastic or neoplastic cell populations. Although the detection of microsatellite alterations does not indicate the specific genetic change in the tumor, detection of clonal cell populations might serve as a cancer screening marker (65). Identical alterations have been found in lung cancers and corresponding sputum samples demonstrating minimal atypia (68). The *p16* gene is located on the short arm of chromosome 9(9p21) and is frequently mutated or inactivated in tumors and cell lines derived from lung cancer (69, 70). Belinsky *et al.* (71) measured hypermethylation of the CpG islands in the sputum of lung cancer patients and demonstrated a high correlation with early stages of non-small cell lung cancer, which indicated that p16 CpG hypermethylation could be useful in the prediction of future lung cancer. However, prospective studies are needed to evaluate the role of p16 hypermethylation as a marker for early lung cancer detection. Multiple other genes are inactivated by hypermethylation in lung cancer (72), and the detection of hypermethylation may be useful for risk assessment and early diagnosis.

Computer-assisted Image Analysis. Computer-assisted image analysis was initially used to detect malignancy-associated changes (*e.g.*, subvisual or nonobvious changes in the

distribution of DNA in the nuclei of histologically normal cells in the vicinity of preinvasive or invasive cancer; 73). In a retrospective analysis of sputum cytology slides, malignancy-associated changes alone correctly identified 74% of the subjects who later developed squamous cell carcinoma (74). The technique has been improved, and recent data showed sensitivities of 75% for stage 0/I lung cancer and 85% for adenocarcinomas with a specificity of 90% (75). This quantitative microscopy technique allows the examining of thousands of cells per slide within a relative short time. Similar techniques have been approved in the United States for cervical cancer screening, and might, in the future, play a role for lung cancer screening. However, no prospective clinical studies has evaluated this technology in a larger lung cancer screening setting.

High Throughput Technology. With future advances in gene chip technology, it might become feasible to probe for expression of multiple genes in sputum specimens of asymptomatic individuals. However, this requires a large amount of undegraded RNA from respiratory tract cells. With the high throughput technology, a higher sensitivity might be achieved by using multiple markers at the cost of achieving a lower specificity, which would be undesirable for a screening study.

In conclusion, we need to reevaluate the role of sputum cytology for screening and early detection of lung cancer because of advances in biomarkers and technology. Ongoing studies with standard and biomarker analysis in high-risk groups might change the previous negative attitude and provide a new perspective on sputum cytology as a mass screening tool when applied in a high-risk population. Adding different molecular diagnostic tests gives the possibility for early diagnosis far in advance of clinical presentation. However, validation of the tests in larger prospective studies is necessary, and the individual tests have to be compared with each other to define the role of early diagnosis in the overall management of high-risk subjects. Furthermore, health economic issues have to be considered.

Specimens for Clinical Testing: BAL

BAL involves the infusion and reaspiration of a sterile saline solution in distal segments of the lung via a fiberoptic bronchoscope. Ahrendt *et al.*, (76) examined a series of 50 resected non-SCLC tumor patients and compared the tumor and BAL with regard to molecular markers including p53 mutations, K-ras mutation, the methylation status of the CpG island of the *p16* gene, and microsatellite alteration (Tables 1 and 2). With the possible exception of the test for microsatellite alteration, all of the tests had relatively high sensitivity and could detect mutant cells in the presence of a large excess of normal cells. The frequencies of these changes in the tumors ranged from 27% (for K-ras mutations) to 56% (for p53 mutations). As expected, p53 mutations were more frequent in central (predominantly squamous cell) tumors, and K-ras mutations were more frequent in peripheral (predominantly adenocarcinoma) tumors. The specificity was high (nearly 100%) because, with the exception of microsatellite alterations, the same genetic change in BAL sample as in tumors was always found, but the sensitivity was low, and in only 53% of tumors that contained molecular lesions were the same abnormalities detected in corresponding BAL fluids. Specifically, the tests were least helpful in the group of patients in whom improved diagnostic abilities are most needed, those with small, peripherally located tumors (77). Unfortunately, the investigators were not able to compare the molecular tests with routine cytopathological analysis of the BAL specimens. The sensitivity of the molecular tests in BAL specimens has to be improved, and we need to know the results from subjects at increased risk (current and former smokers without lung cancer and survivors of previous cancer of the upper respiratory tract) and subjects with chronic lung diseases as well as results from healthy never smokers.

A European group has previously shown that genetic alterations detected in DNA from bronchial lavage of individuals with lung cancer were also found in individuals with no evidence of malignant disease (78), which raises the question about the specificity of such molecular damage in neoplastic conditions. To improve the sensitivity and specificity of detecting allelic imbalance in lung tumors, high throughput PCR-based microsatellite assays have been established (79). In a recent study by Fielding *et al.* (80), the up-regulation of hnRNP A2/B1 was found to be a promising marker in BAL for the detection of premalignant and malignant bronchial lesions with a diagnostic sensitivity of 96% and a specificity of 82%.

It is too early yet to make conclusions as to whether BAL examinations will add to other pathological/molecular biological clinical studies. To obtain diagnostic material for BAL bronchoscopy is required, and we do not have any data that compare BAL examinations with biopsies. Thus, we do not know whether BAL is a valuable adjunct to the biopsies taken under the same bronchoscopy procedure.

Specimens for Clinical Testing: Peripheral Blood

For many years scientists have searched for a lung cancer-specific tumor marker that could be detected in peripheral blood. Optimism was raised in the "early" immunocytochemistry era by the use of monoclonal antibodies raised against more-or-less specific epithelial epitopes. In the search for epithelial cells in peripheral blood and bone marrow, monoclonal antibodies against cytokeratin have been used. However, these reactions are clearly not cancer-specific, and some antibodies have been shown to cross-react with normal blood or bone marrow elements (81, 82). Another explanation could be that cells from the macrophage/monocyte system may contain proteins derived from the primary tumor that have undergone necrosis and apoptosis and that these processed proteins are recognized by the antibodies (82). On the basis of "traditional" immunocytochemistry, no markers have been able to detect premalignant or early-malignant disorders based on a peripheral blood sample. However, with the development of DNA technologies, new possibilities have been raised, and, with the use of PCR techniques, some promising reports have been published.

Nanogram quantities of DNA circulating in blood are present in healthy individuals (83, 84). Tumor DNA is also released into the plasma component in increased quantities (85, 86). Thus, the plasma and serum of cancer patients is enriched in DNA, an average four times the amount of free DNA as compared with normal controls (87). In a study by Chen *et al.* (88), a comparison of microsatellite alterations in tumor and plasma DNA was done in SCLC patients, and 93% of the patients with

Table 1 Tissues and other resources for the study of molecular markers

Specimen	Ref.	Comments
Tumor tissue	Numerous	Mixture of cell types, may require microdissection (139). Extensively used for most studies. Alcohol-fixed fine-needle aspirates may be used for mutational and other studies.
Sputum	65, 68, 71, 74	Respiratory cells usually in small minority. Most samples fixed in Saccomanno's fixative. Several studies have been performed on these specimens many years later.
Surrogate organ	140	Predominantly squamous epithelial cells. Buccal smears, brushings of tongue or tonsil may be explored as potential surrogate organs resulting from the field effect of tobacco damage of the entire upper aerodigestive tract. This concept needs to be confirmed.
Bronchial brush/wash	141, 142, 143	Predominantly respiratory cells. Fresh, frozen, or alcohol-fixed samples are suitable for multiple studies including FISH.[a]
Bronchial tissues	42, 43, 45, 144, 145	Usually from bronchial biopsies, but may be obtained from surgical resection specimens. Formalin fixation and paraffin embedding required for histological diagnosis, although EASI preps may permit identification and isolation of subpopulations. Paraffin sections may be used for genotyping polymorphisms, for allelotyping, and for *in situ* hybridization.
BAL fluids	76, 78, 146, 147, 148	BAL fluids are useful for examining the peripheral airway cells, which are the precursor cells of most adenocarcinomas. Numerous mononuclear cells present. Enrichment of epithelial cells desirable.
Blood components	72, 92, 149	Analysis of circulating tumor cells and genetic material shed by dying tumor cells into the plasma component may yield useful biological and diagnostic information. Gene mutations and presence of epithelial cell markers have been used to detect circulating tumor cells. Gene mutations, allelic loss, microsatellite alterations, and aberrant methylation have been used to identify tumor cell DNA released into the fluid compartment.
Tissue for molecular staging	150, 151	Although little data exists for lung cancers, regional lymph nodes, sentinel lymph nodes, and surgical resection margins have been used in other tumor types for molecular staging.
Tumor cell lines	152, 153	Provide an unlimited self-replicating source of high-quality molecular reagents and have been used for numerous studies. Cell lines may or may not reflect the properties of the tumors from which they were derived (26), although they probably represent cellular subpopulations (27). Aggressive metastatic tumors are more likely to be successfully cultured (28) resulting in skewed data.
Cultures of nonmalignant tissues	154, 155	Epithelial cultures may be useful for studying molecular changes during multistage pathogenesis. Temporary as well as a few immortalized cultures from nonmalignant epithelial cells have been established. B-lymphoblastoid cultures are useful for linkage analysis, for genetic susceptibility studies, and for allelotyping corresponding tumors.
Nonmalignant tissue from patients and from cancer-free relatives	156, 157, 158	Tissues such as buccal smears, tumor-free lymph nodes, and peripheral blood mononuclear cells are useful as controls for linkage analysis, for genetic susceptibility studies, and for allelotyping corresponding tumors.

[a] FISH, fluorescence *in situ* hybridization; EASI, epithelial aggregate separation and isolation.

microsatellite alterations in tumor DNA also had modifications in the plasma DNA. However, some patients had LOH only in the tumor DNA. Because most of the microsatellite alterations were similar in tumor DNA and plasma DNA, they concluded that some of the DNA circulating in the blood comes from the tumor. Thus, modifications of circulating DNA can be used as an early detection marker. Detection of aberrant DNA methylation in serum DNA in patients with non-SCLC has been reported (72). Although the number of patients was small and the hypermethylated DNA was found in all stages, it opens up for the possibility to be used as an early lung cancer detection marker. Furthermore, *p53* and *ras* gene mutations have been detected in the plasma and serum of patients with colorectal cancers (89–91), pancreatic carcinomas (92, 93), and hematological malignancies (94).

In conclusion, the limited direct accessibility of lung carcinomas has led to efforts to identify tumor-associated soluble markers in serum or plasma. Many of the currently recognized soluble markers were first identified as "tumor" markers but, when evaluated in nonneoplastic tissue, have often been found in normal cells as well as in tumors. For early detection of lung cancer, we need more clinical data evaluating these new molecular biological markers from multiple sites, especially in high-risk groups.

Table 2 Molecular approaches for lung cancer investigation

Specimen	Ref.	Comment
Gene mutations	159, 160, 161	Widely used technique, especially for *p53* and *ras* genes. Often used to determine the rôle of a newly discovered gene in the pathogenesis of lung cancer. May be of diagnostic and prognostic significance. Multiple methodologies available.
Allelotyping	18, 158	Useful as a partial substitute for mutational analysis and for determining the chromosomal locations of putative tumor suppressor genes. Widely used to study multistage pathogenesis. Readily performed on formalin-fixed and microdissected tissues. Increasing use of genotyping using automatic sequencers.
Gene expression at RNA and protein level	145, 162, 163, 164, 165, 166	Northern blotting and reverse transcription-PCR are widely used to investigate gene expression. Western blotting often used for detection of protein expression. *In situ* hybridization for message expression can be performed on paraffin-embedded tissues and, thus, can be used to investigate multistage pathogenesis. Microarray techniques offer promise of examining all or most of the genome but currently require relatively large amount of high-quality RNA from purified cell populations. Sage technique useful for investigation and identification of expressed genes. Similarly, advances in proteomics will permit simultaneous detection of multiple proteins. Numerous immunohistochemical studies of oncogene expression have been used to study multistage pathogenesis. Of particular interest, hnRNP expression on exfoliated epithelial cells in sputum samples may predict for development of cancer.
Molecular cytogenetics	40, 167, 168, 169, 170	*In situ* hybridization studies of fixed materials or using smears has provided considerable information about numerical and structural changes.
Comparative genomic hybridization	171, 172	Useful for detection of gene amplifications. Less sensitive for the detection of regions of allelic loss.
Morphometric studies	74, 173, 174	May be applied to paraffin-embedded tissues. Useful for determining aneuploidy and for measuring a number of nuclear and cytoplasmic parameters.

Specimens for Clinical Testing: Bronchoscopy

WLB is the most commonly used diagnostic tool for obtaining a definite histological diagnosis of lung cancer. Bronchoscopy has major diagnostic limitations for premalignant lesions. Because these lesions are only a few cells thick (0.2–1 mm) and have a surface diameter of only a few millimeters, they rarely are observed as visual abnormalities. Woolner (95) reported that squamous cell CIS was visible to experienced bronchoscopists in only 29% of cases. To address this limitation, fluorescence bronchoscopy was developed. Early studies of fluorescence bronchoscopy entailed the use of fluorescent drugs (hematoporphyrin dyes) that were preferentially retained in malignant tissue (96). Although, studies evaluating this approach did, in fact, show that early invasive and *in situ* cancers could be localized, the detection of dysplasia remained problematic (97–100). Furthermore, the development of photodynamic diagnostic systems was hampered by problems including skin photosensitizing and interference with tissue autofluorescence. To overcome these problems, a new laser photodynamic diagnostic system was developed (101). This system detected tumor-specific drug fluorescence at 630 nm wavelength, which is far from normal tissue autofluorescence (500–580 nm), and interference by autofluorescence from normal tissue should then have been eliminated, but it remained a significant problem (102).

Another approach was developed by Palcic *et al.* (103), who noticed the lack of autofluorescence in the tumor lesions by using blue light (442 nm) rather than white light to illuminate the bronchial surface. They amplified the difference in autofluorescence between normal, premalignant, and tumor tissue for clinical use (103, 104). Using a high-quality-charge coupled device and special algorithm, the LIFE was developed, taking advantage of the principle that dysplastic and malignant tissues reduce autofluorescent signals compared with normal tissue (Fig. 3).

Several studies have been performed comparing the diagnostic specificity and sensitivity of LIFE bronchoscopy *versus* WLB in diagnosing preinvasive and early-invasive lesions (105–109; Table 3). Most of the studies reported a higher diagnostic sensitivity of LIFE bronchoscopy in the detection premalignant and early-malignant lesions at the cost of lower specificity (*i.e.*, more false-positive results). In most of these studies, lesions with moderate dysplasia or worse were the target of the study and rated as "positive." The prevalence of preinvasive and early lung cancer varies widely from one study to another, from 20.2% (105) to 65.8% (102). The explanation might be beyond the risk profile of genetic variations or different levels of experience among the endoscopists as well as the pathologists involved. Furthermore, there seems to be a training effect in using the LIFE bronchoscope, which has been demon-

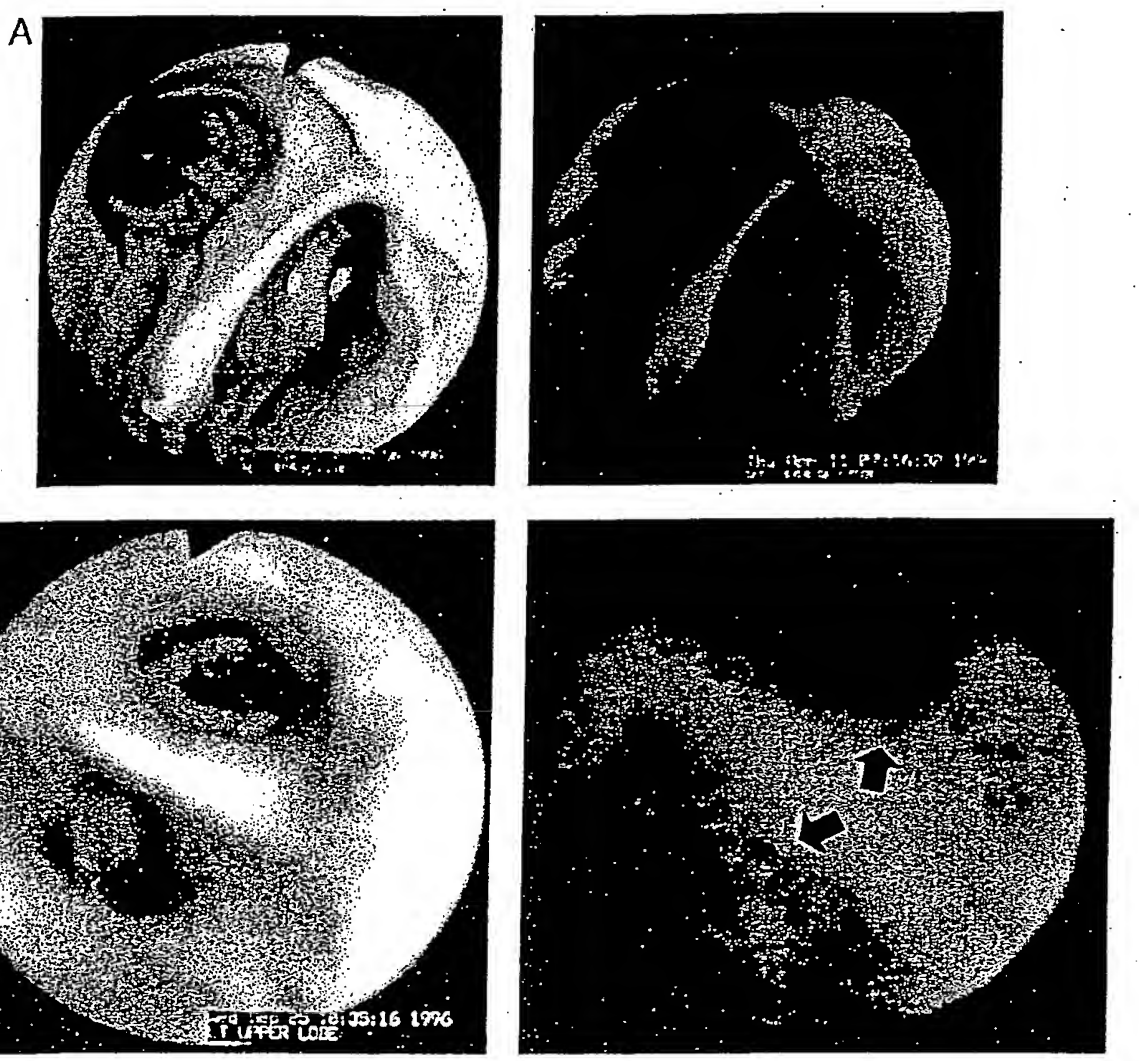


Fig. 3 *A*, normal WLB and normal LIFE bronchoscopy. *B*, WLB shows inflammatory changes in the bronchial mucosa but no suspicion of malignancy (*left*). LIFE bronchoscopy shows diffuse reduced autofluorescence (visualized by *diffuse red-brownish colorization*; *arrows*). Biopsy demonstrated diffuse severe dysplasia.

strated by Venmans *et al.* (107). In their study, the diagnostic sensitivity increased from 67 to 80% when comparing the first and the second half of the study. The use of the LIFE device in conjunction with WLB improved the detection rate of preneoplastic lesions and CIS significantly (Table 3). Kurie *et al.* (106) looked for more subtle tissue transformation, but their study included few patients with moderate dysplasia or worse. No improvement in the evaluation of metaplasia index was observed by the use of LIFE bronchoscopy. Thus, differences in the study population might explain the different conclusion. There are still no clinical studies with sufficient long-term data showing that moderate dysplasia is the most relevant clinical predictor of eventual malignancy. Limitations in making conclusions from the existing studies are also the potential methodological bias related to the order in which the different bronchoscopy procedures are done and whether the same examiner has performed both procedures. To address these issues, a prospective randomized study between LIFE bronchoscopy and WLB was done at the University of Colorado Cancer Center. The study design included a randomization with regard to the order of procedure as well as the order of the individual bronchoscopist (109). The order of the procedure and of the individual bronchoscopist did not affect the results. The study also demonstrated a significantly higher sensitivity in detecting premalignant lesions visualized by the LIFE, but at the cost of a lower specificity (109). The reason for the low diagnostic specificity found with the LIFE bronchoscopy in the different studies might be attributable to the visualization of more abnormal foci with the LIFE bronchoscope, with the consequence that a larger number of biopsies were taken and, thus, there was a higher risk of more false-positive results. The use of LIFE bronchoscopy has led to the identification of a new morphological entity, the ASD, which is described above. In a recent morphological study angiodysplastic changes were frequently found in preneoplastic

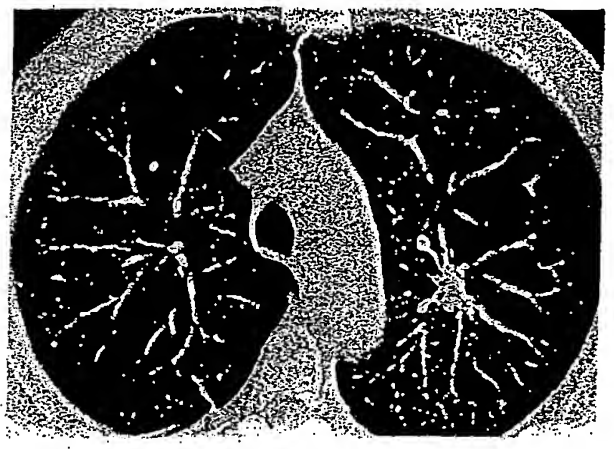


Fig. 4 Seventy-one-year-old man with a spicular nodule in upper left lobe demonstrated on low-dose helical CT (*picture*), but not visible on chest X-radiography. CT- guided biopsy showed adenocarcinoma.

and early-malignant lesions in the bronchi (26). The morphological entity has been confirmed in preneoplasias among smokers, and the perspectives of this finding have been extensively discussed (110). The prognostic significance of this morphological entity is currently studied in ongoing long-term follow-up studies. Future studies have to evaluate the role of ASD as a biomarker for early lesions and whether it can be used as a marker for treatment effect or therapeutic target for chemoprevention.

The LIFE bronchoscope may play an important role in the screening and follow-up of subjects at high risk of developing lung cancer. At this stage, however, it is unknown whether the LIFE bronchoscope will lead to a reduction in lung cancer mortality. There are also no data on cost-effectiveness and cost-benefit analyses available for this new diagnostic procedure. The use of the LIFE bronchoscope may also in the future be extended to other indications, *e g.*, patients staged as having resectable lung cancer on one side. Whether LIFE bronchoscopy of the contralateral lung will disclose abnormalities, which would change the therapeutic decision, is not yet reported.

Recent Advances in Radiology

The previous NCI-sponsored screening trials failed to demonstrate any reduction in the lung cancer mortality by sputum cytology and yearly chest radiography as mass screening tools for lung cancer screening. Limitations of design and execution of the studies, however, have been discussed extensively (8, 111, 112). An extended follow-up (median, 20.5 years) of the Mayo Lung Project was recently published (113). There was still no difference in lung cancer mortality between the intervention arm and the control arm (4.4 *versus* 3.9 deaths per 1000 person-years). However, the median survival for patients with resected early-stage disease was 16.0 years in the intervention arm *versus* 5.0 years in the usual-care arm ($P < 0.05$). The latter findings have raised the question as to whether some small lesions with limited clinical relevance may have been identified in the intervention arm, and the question of "overdiagnosis" was discussed in accompanying editorials (114).

Mass screening for lung cancer has been performed in Japan for many years and has been performed in over 500,000 people in about 80% of the local communities (115). Sobue *et al.* (116) observed that annual clinic-based chest X-ray screening for lung cancer in Japan showed reduced lung cancer mortality by about one-fourth among individuals who underwent screening once a year. In this screening program, the relative odds ratio of dying from lung cancer within 12 months was 0.535 and in the 12–24-month period was 0.638 (117). However, many studies have focused on the pitfalls in the detection of abnormalities by radiography (118–122). The limit of chest radiographic sensitivity for nodule detection is roughly 1 cm in diameter, by which time the tumor has over 10^9 cells and may already have violated bronchial epithelium and vascular epithelium. CT has been shown to be more effective in the detection of peripheral lung lesions compared with plain radiography or conventional tomography of the whole lung (123, 124).

Spiral CT scan is a relatively new technology with the ability to continuously acquire data resulting in a shorter scanning time, a lower radiation exposure, and improved diagnostic accuracy compared with those of plain radiography (125–127). Spiral CT allows the whole chest to be imaged in one or two breath-holds, reducing motion artifacts and eliminating respiratory misregistration or missing nodules. Although there is greater radiation exposure with CT than with chest radiography, low-dose techniques (lower mA of 30–50 compared with 200 for conventional CT) have achieved calculated exposure doses that are 17% that of conventional CT and 10 times that of chest radiographs. Further reduction in radiation dose while maintaining diagnostic accuracy is a topic of current research. Furthermore, for the baseline screening, low-dose spiral-CT-scan i.v. contrast is not administered. Nodules as small as 1–5 mm can be shown with modern spiral CT technology (25, 128). The obvious advantages with this new technology led some groups in Japan and in the United States to look to low-dose spiral CT as a tool for screening (Refs. 129–131; Tables 4 and 5).

In a Japanese report, spiral CT scans and chest radiographs were done twice a year in 1369 individuals (129). Peripheral lung cancer was detected in 15 (0.3%) of 3457 examinations, and, among the 15 lung cancer cases detected, the results of chest X-ray were negative in 11(73%), and the tumors were detected only by low-dose spiral CT. The detection rates of low-dose spiral CT and chest X-ray were 0.43% (15 of 3457 examinations) and 0.12% (4 of 3457 examinations), respectively. Furthermore, 14 (93%) of the 15 lung cancers were stage I disease. The histology showed that 11 of the 15 lung cancer cases were adenocarcinoma, and 4 had squamous cell carcinoma. The effective exposure dose with spiral CT scan in that study was calculated to about one-sixth that of conventional CT.

The ELCAP in New York was designed to determine: (*a*) the frequency with which nodules were detected; (*b*) the frequency with which detected nodules represent malignant disease; and (*c*) the frequency with which malignant nodules are curable (131). In the ELCAP study, 27 lung cancers were found among 1000 subjects screened. Among the 27 patients with cancer, 85% had stage I disease (Table 5).

Another population-based study on low-dose CT screening has been published by Sone *et al.* (130), using a mobile low-dose spiral CT scanner. The detection rate was 0.48% (*i.e.*, 4–5

Table 3 Bronchoscopy *versus* WLB in diagnosing premalignant and early-malignant lesions

Author	No. of biopsies	Sensitivity LIFE+WLB	Sensitivity LIFE	Sensitivity WLB	Relative sensitivity LIFE+WLB	Specificity LIFE+WLB	Specificity LIFE	Specificity WLB	Relative specificity LIFE+WLB	Predictive values PPV[a] LIFE+WLB	Predictive values NPV LIFE+WLB	PPV LIFE	NPV LIFE	PPV WLB	NPV WLB
Lam *et al.* (105)	700	0.67	NR	0.25	6.3 (2.7)[c]	0.66	NR	0.90	NR	0.33	0.89	NR	NR	0.39	0.83
Kurie *et al.*[b] (106)	234	NR	0.38	NR	NR	NR	0.56	NR	NR	NR	NR	0.16	0.81	NR	NR
Venmans *et al.* (107)	139	NR	0.89	0.78	1.43	NR	0.61	0.88	NR	0.20	NR	0.14	0.99	0.32	0.98
Vermulen *et al.* (108)	172	0.93	NR	0.25	3.75	0.21	NR	0.87	NR	0.13	0.96	NR	NR	0.19	0.90
Kennedy *et al.* (109)	394	0.79	0.72	0.18	4.4	0.3	0.43	0.78	0.38	0.21	0.85	0.25	0.87	0.17	0.80

[a] PPV, positive predictive value; NPV, negative predictive value; NR, not reported.
[b] Based on reference pathologist.
[c] If invasive carcinoma is included.

Table 4 Results from three population-based screening studies with low-dose spiral CT (LDCT)

Authors	No. of individuals studied	True positive *n*	False positive[a] %	Predictive value %	Detection rate % LDCT	Detection rate % X-ray	Pack-yr	Age incl. yr
Kaneko *et al.* (129)	1369	15	15.6	6.6	0.43	0.12	>20	>50
Sone *et al.* (130)	3967	19	5.0	8.8	0.46–0.5		>30[b]	40–74
Henschke *et al.* (131)	1000	27	20.1	11.6	2.7	0.70	>10[c]	>60

[a] Defined as individuals with "test-positive," in whom further workup gave no suspicion of malignancy.
[b] The study also included a group of nonsmokers.
[c] Average = 45 (not reported in the other studies).

Table 5 Histology, stage, and size of primary lung cancer detected by low-dose spiral CT

Author	No. of cancers/ No. screened	Histology % Adeno[a]	Squam.	Other	TNM % I	II	III	IV	Size (mm) Average	Range	<10	11–20	>21
Kaneko *et al.* (129)	15/1369 (1.1%)	73	17		93		7		12	8–18			
Sone *et al.* (130)	19/5483 (0.3%)	63	5	32	84			16	17	6–47	4	14	3
Henschke *et al.* (131)	27/1000 (2.7%)	67	3	30	85	4	11				15	8	4

[a] Adeno, adenocarcinoma; Squam., squamous cell carcinoma; TNM, tumor-node-metastasis.

cases per 1000 examinations). Surprisingly, there was no difference in the detection rate among smokers (0.52%) *versus* nonsmokers (0.46%). The results from the three population-based studies are summarized in Tables 4 and 5. The conclusion from these studies is that 85% of the lung cancers detected by low-dose CT were in stage I, offering improved possibility for curative treatment and better prognosis in general. However, the issue of "false-positive" scans has to be taken into consideration. Thus far, up to 20% of the participants with nodules on the scan had no malignancy during the follow-up period. The possibility that the cancers found represent incidental cancers as in the Mayo Lung Project must also be considered (114). The results from these studies confirm the expectation that low-dose CT increases the detection of small noncalcified nodules and, that lung cancer at an earlier and more curable stage are detected. The mobile CT screening study by Sone *et al.* (130) showed that low-dose CT increased the likelihood of detection of malignant disease 10 times as compared with radiography. The overall rate of malignant disease was lower in the Japanese studies (129, 130) compared with the ELCAP study (Ref. 131; detection rates 0.43–0.48% *versus* 2.7%). This could be because the Japanese studies screened individuals from the general population ages 40–74, whereas ELCAP screened people at high risk, ages ≥60, with a tobacco history of at least 10 pack-years. Thus, as expected, the risk of the population to be screened affects the rate of cancer detection.

Questions remaining to be answered include: (*a*) what are the diagnostic sensitivity and specificity of this procedure; and (*b*) does screening reduce lung cancer mortality? The spiral CT has not been as sensitive for small central cancers as it is for small peripheral cancers (129, 131). Minute nodules of lung cancer that are near the threshold of detectability may be overlooked at spiral CT screening (132). A prospective study of the diagnostic sensitivity of spiral CT has recently shown that the diagnostic sensitivity exceeded the sensitivity of conventional CT in previous reports (25). However, there were limitations in the detection of intrapulmonary nodules smaller than 6 mm and of pleural lesions. Compared with surgery (thoracotomy with palpation of deflated lung, resection, and histology), the sensitivity of spiral CT was 60% for intrapulmonary nodules of <6 mm and 95% for nodules of ≥6 mm and was 100% for neoplastic lesions ≥6 mm. Furthermore, a marked difference in the sensitivities of two independent observers was found for nodules smaller than 6 mm, whereas agreement was much better for

6–10-mm nodules (25). Given these promising preliminary clinical results, further research is needed to determine the optimal technique for spiral CT screening, which includes collimation, reconstruction interval, pitch, and viewing methods. Decreasing the slice thickness to 3 mm, monitoring the viewing of examinations, and computer-aided diagnosis have been used to improve the diagnostic capability of spiral CT in the detection of pulmonary nodules (133–136).

Future large scale randomized studies have to confirm whether in fact spiral CT screening will lead to a reduction in lung cancer mortality. In a randomized study, the following questions arise: (*a*) what is the optimal high-risk group to study and what should be the control arm? (*b*) what should be the end points (goals) of the studies? The ultimate goal is to reduce the lung cancer mortality. However, although this is a long-term goal, intermediate end points from such studies should be evaluated. The change to more curable stages at diagnosis for the lung cancer patients is one such immediate goal; (*c*) what is the optimal workup and the morbidity of this program? (*d*) what is the cost of such a screening program? and (*e*) what is the false-positive rate of the screening findings? Incorporation of smoking cessation programs should be included in the future design of screening studies because it has been shown that screening with low-dose CT in participants who are still smoking provides substantial motivation for smoking cessation (137).

The studies with spiral CT-scan have demonstrated the superior diagnostic ability in the detection of small peripherally located tumors, most of the malignant ones of adenocarcinoma type of histology. The diagnostic sensitivity of spiral CT for more centrally located tumors (mostly squamous cell carcinoma) is significantly lower than for the peripherally located ones. Through these spiral CT studies, we will learn about the biology, pathology, and clinical course of these small tumors, which might be different from what we know about clinically more evident tumors detected routinely in previous studies.

Because lung cancer is so common, the introduction of any new screening technique in this area has to be underpinned by careful definition of the cost implications and must be justified by compelling evidence. The cost-effectiveness of the spiral CT approach should be assessed by evaluating the rate of over-diagnosing nonmalignant, relatively common abnormalities and comparing CT imaging to other diagnostic technologies.

PET with FDG has recently emerged as a practical and useful imaging modality in the preoperative staging of patients with lung cancer. However, whereas CT is most frequently used to provide additional anatomical and morphological information about lesions, the FDG PET imaging provides physiological and metabolic information that characterizes lesions that are indeterminate by CT. FDG PET imaging takes advantage of the increased accumulation of FDG in transformed cells and is sensitive (~95%) for the detection of cancer in patients who have indeterminate lesions on CT (138). The specificity (~85%) of PET imaging is slightly less than its sensitivity because some inflammatory processes avidly accumulate FDG. The high negative predictive value of PET suggests that lesions considered negative on the study are benign, biopsy is not needed, and radiographic follow-up is recommended. Several studies have documented the increased accuracy of PET compared with CT in the evaluation of the hilar and mediastinal lymph node status in patients with lung cancer (138). However, the PET resolution is sufficient only for nodules ≥6 cm and will not be helpful in detecting the very small nodules. Compared with low-dose spiral CT, the FDG PET scan is more expensive and time consuming. The role of PET scan in early diagnosis of lung cancer in an asymptomatic high-risk population is not yet evaluated. However, future studies have to include PET evaluation to define its role in a population screening setting.

Conclusion

Recent advances in molecular biology and pathology have led to a better understanding and documentation of morphological changes in the bronchial epithelium before development of clinical evident lung carcinomas. Combined with technical developments in radiological and bronchoscopic techniques, these procedures offer great promise in diagnosing lung cancer far in advance of clinical presentation. Any of these individual procedures could be incorporated into the routine management of individuals at risk for developing primary or secondary lung cancer, and for several of these methods, clinical studies are under way. Preliminary reported data are very promising for the early detection of lung cancer. Future studies must incorporate the different methods in a multidisciplinary scientific setting to evaluate the role of the individual method in the overall management for individuals at high risk for developing lung cancer. Several of these tests might diagnose the disease at the stage of clonal expansion before invasive carcinoma has developed. A management and intervention strategy appropriate to that stage of disease have to be developed. Preliminary studies of chemoprevention agents are reported, and new agents based on other biological mechanisms are under development and ready for clinical trials. It is now time to plan clinical trials that evaluate both diagnostic and therapeutic approaches to access their impact on the incidence of clinical lung cancer.

Acknowledgments

We thank Drs. Stephen Lam, Vancouver, British Columbia, Canada, and Kavita Garg, University of Colorado Health Sciences Center, Denver, Colorado, for a critical review of the manuscript and Drs. Timothy Kennedy and York Miller for submitting illustrations for white-light and LIFE bronchoscopy.

References

1. Greenlee, R. T., Murray, T., Bolden, S., and Wingo, P. A. Cancer Statistics, 2000. CA Cancer J. Clin., *50:* 7–30, 2000.

2. Mountain, C. F. Revision in the international system for staging of lung cancer. Chest, *111:* 1710–1717, 1997.

3. Ihde, D. C. Chemotherapy of lung cancer. N. Engl. J. Med., *327:* 1434–1441, 1992.

4. Melamed, M. R., Flehinger, B. J., Zaman, M. B., Heelan, R. T., Perchick, W. A., and Martini, N. Screening for lung cancer: results of the Memorial Sloan-Kettering study in New York. Chest, *86:* 44–53, 1984.

5. Fontana, R. S., Sanderson, D. R., Woolner, L. B., Taylor, W. F., Miller, W. E., and Muhm, J. R. Lung Cancer Screening: The Mayo program. J. Occup. Med., *28:* 746–750, 1986.

6. Tockman, M. S. Survival and mortality from lung cancer in a screened population: The Johns Hopkins Study. Chest, *89:* 324S–325S, 1986.

7. Kubik, A., Parkin, D. M., Khlat, M., Erban, J., Polak, J., and Adfamec, M. Lack of benefit from semi-annual screening for cancer of the lung: follow-up report of a randomized controlled trial on a population of high-risk males in Czechoslovakia. Int. J. Cancer, *45:* 26–33, 1990.

8. Fontana, R. S., Sanderson, D. R., Woolner, L. B., Taylor, W. F., Miller, W. E., Muhm, J. R., Bernatz, P. E., Payne, W. S., Pairolero, P. C., and Bergstralh, E. J. Screening of lung cancer. A critique of the Mayo Lung Project. Cancer (Phila.), *67:* 1155–1164, 1991.

9. Strauss, G. M., Gleason, R. E., and Sugarbaker, D. J. Screening for lung cancer. Another look: a different view. Chest, *111:* 754–768, 1997.

10. Hirsch, F. R., Brambilla, E., Gray, N., Gritz, E., Kelloff, G. J., Linnoila, I., Pastorino, U., and Mulshine, J. L. Prevention and early detection of lung cancer—clinical aspects. Lung Cancer (Limerick), *17:* 163–174, 1997.

11. Hong, W. K. Chemoprevention of Lung Cancer. Oncology (Basel), *13* (Suppl. 5): 135–141, 1999.

12. Mulshine, J. L., and Henschke, C. I. Prospects for lung-cancer screening. Lancet, *355:* 592–593, 2000.

13. Travis, W. D., Colby, T. V., Corrin, B., Shimosato, Y., and Brambilla, E. Histological typing of tumours of lung and pleura. *In:* L. H. Sobin (ed.), World Health Organization International Classification of Tumours, Ed. 3. New York: Springer-Verlag, 1999.

14. Franklin, W. A. Pathology of lung cancer. J. Thorac. Imaging, *15:* 3–12, 2000.

15. Colby, T. V. Precursor lesions to pulmonary neoplasia. *In:* C. Brambilla and E. Brambilla (eds.), Lung Tumors. Fundamental Biology and Clinical Management, pp. 61–87. New York: Marcel Dekker Inc.; 1999.

16. Peters, E. J., Morice, R., Benner, S. E., Lippman, S., Lukeman, J., Lee, J. S., Ro, J. L., and Hong, W. K. Squamous metaplasia of the bronchial mucosa and its relationship to smoking. Chest, *103:* 1429–1432, 1993.

17. Auerbach, O., Gere, B., Forman, J. B., Petrick, T. G., Smolin, H. J., Muchsam, G. E., Kassouny, D. Y., and Stout, A. P. Changes in bronchial epithelium in relation to smoking and cancer of the lung. N. Engl. J. Med., *256:* 97–104, 1957.

18. Wistuba, I. I., Behrens C. Milchgrub, S., Bryant, D., Hung, J., Minna, J. D., and Gazdar, A. F. Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. Oncogene, *18:* 643–650, 1999.

19. Lam, S., LeRiche, J. C., Zheng, Y., Coldman, A., MacAulay, C., Hawk, E., Kelloff, G., and Gazdar, A. F. Sex-related differences in bronchial epithelial changes associated with tobacco smoking. J. Natl. Cancer Inst., *91:* 691–696, 1999.

20. Saccomanno G. Archer, V. E., Auerbach, O., Saunders, R. P., and Brennan, L. M. Development of carcinoma of the lung as reflected in exfoliated cells. Cancer (Phila.), *32:* 256–270, 1974.

21. Frost, J. K., Ball, W. C., Jr., Levin, M. L., Tockman, M. S., Erozan, Y. S., Gupta, K., Eggleston, J. C., Pressman, N. J., Donithan, M. P., and Kimball, A. W. Sputum cytopathology: use and potential in monitoring the workplace environment by screening for biological effects of exposure. J. Occup. Med., *28:* 692–703, 1986.

22. Venman, B. J. W., van Boxem, T. J. M., Smit, E. F., Postmus, P. E., and Sutedja, T. G. Outcome of bronchial carcinoma *in situ*. Chest, *117:* 1572–1576, 2000.

23. Slebos, R. J., Baas, I. O., Clement, M. J., Offerhaus, G. J., Askin, F. B., Hruban, R. H., and Westra, W. H. p53 alterations in atypical alveolar hyperplasia of the human lung. Hum. Pathol., *29:* 801–808, 1998.

24. Kitamura, H., Kameda, Y., Ito, T., and Hayashi, H. Atypical adenomatous hyperplasia of the lung: Implications for the pathogenesis of peripheral lung adenocarcinoma. Am. J. Clin. Pathol., *111:* 610–622, 1999.

25. Diederich, S., Semik, M., Lentschig, M. G., Winter, F., Scheld, H. H., Roos, N., and Bongartz, G. Helical CT of pulmonary nodules in patients with extrathoracic malignancy: CT-surgical correlation. Am. J. Roentgenol., *172:* 353–360, 1999.

26. Keith, R. L., Miller, Y. E., Gemmill, R. M., Drabkin, H. A., Dempsey, E. C., Kennedy, T. C., Prindiville, S., and Franklin, W. A. Angiogenic squamous dysplasia in bronchi of individuals at high risk for lung cancer. Clin. Cancer Res., *6:* 1616–1625, 2000.

27. Muller, K. M., and Muller, G. The ultrastructure of preneoplastic changes in bronchial mucosa. Curr. Top. Pathol., *73:* 233–263, 1983.

28. Hirsch, F. R., Gazdar, A. F., Gabrielson, E., Lam, S., and Franklin, W. A histopathologic evaluation of premalignant and early malignant bronchial lesions: an interactive program based on internet digital images to improve WHO criteria for early diagnosis of lung cancer and for monitoring chemoprevention studies—a SPORE collaborative project. Lung Cancer (Limerick), *29* (Suppl. 2): 209, 2000.

29. Lengauer, C., Kinzler, K. W., and Vogelstein, B. Genetic instabilities in human cancers. Nature (Lond.), *396:* 643–649, 1998.

30. Fong, K. M., Sekido, Y., and Minna, J. D. Molecular pathogenesis of lung cancer. J. Thorac. Cardiovasc. Surg., *118:* 1136–1152, 1999.

31. Hirano, T., Franzen, B., Kato, H., Ebihara, Y., and Auer, G. Genesis of squamous cell lung carcinoma: sequential changes of proliferation, DNA ploidy, and p53 expression. Am. J. Pathol., *144:* 296–302, 1994.

32. Betticher, D. C., Heighway, J., Thatcher, N., and Hasleton, P. S. Abnormal expression of CCND1 and RB1 in resection margin epithelia of lung cancer patients. Br. J. Cancer, *75:* 1761–1768, 1997.

33. Satoh, Y., Ishikawa, Y., Nakagawa, K., Hirano, T., and Tsuchiya, E. A follow-up study of progression from dysplasia to squamous cell carcinoma with immunohistochemical examination of p53 protein overexpression in the bronchi of ex-chromate workers. Br. J. Cancer, *75:* 678–683, 1997.

34. Li, Z. H., Zheng, J., Weiss, L. M., and Shibata D. c-k *ras*, and p53 mutations occur very early in adenocarcinoma of the lung. Am. J. Pathol., *144:* 303–309, 1994.

35. Brambilla, E., Gazzeri, S., Lantuejoul, S., Coll, J. L., Moro, D., Negoescu, A., and Brambilla, C. p53 mutant immunophenotype and deregulation of p53 transcription pathway (Bcl2, Bax, and Waf1) in precursor bronchial lesions of lung cancer. Clin. Cancer Res., *4:* 1609–1618, 1998.

36. Hung, J., Kishimoto, Y., Sugio, K., Virmani, A., McIntire, D. D., Minna, J. D., and Gazdar, A. F. Allele-specific chromosome 3p deletions occur at an early stage in the pathogenesis of lung carcinoma. J. Am. Med. Assoc., *273:* 558–563, 1995.

37. Kishimoto, Y., Sugio, K., Hung, J. Y., Virmani, A. K., McIntire, D. D., Minna, J. D., and Gazdar, A. F. Allele-specific loss in chromosome 9p loci in preneoplastic lesions accompanying non-small cell lung cancers. J. Natl. Cancer Inst., *87:* 1224–1229, 1995.

38. Sugio, K., Kishimoto, Y., Virmani, A. K., Hung, J. Y., and Gazdar, A. F. K-*ras* mutations are a relatively late event in the pathogenesis of lung carcinomas. Cancer Res., *54:* 5811–5815, 1994.

39. Wistuba, I. I., Behrens, C., Virmani, A. K., Milchgrub, S., Syed, S., Lam, S., Mackay, B., Minna, J. D., and Gazdar, A. F. Allelic losses at chromosome 8p21–23 are early and frequent events in the pathogenesis of lung cancer. Cancer Res., *59:* 1973–1979, 1999.

40. Varella-Garcia, M., Gemmill, R. M., Rabenhorst, S. H., Lotto, A., Drabkin, H. A., Archer, P. A., and Franklin, W. A. Chromosomal duplication accompanies allelic loss in non-small cell lung carcinoma. Cancer Res., *58:* 4701–4707, 1998.

41. Westra, W. H., Baas, I. O., Hruban, R. H., Askin, F. B., Wilson, K., Offerhaus, G. J., Slebos, R. J. K-ras oncogene activation in atypical alveolar hyperplasias of the human lung. Cancer Res., *56:* 2224–2228, 1996.

42. Wistuba, I. I., Lam, S., Behrens, C., Virmani, A. K., Fong, K. M., LeRiche, J., Samet, J. M., Srivastava, S., Minna, J. D., and Gazdar, A. F. Molecular damage in the bronchial epithelium of current and former smokers. J. Natl. Cancer Inst., *89:* 1366–1377, 1997.

43. Mao, L., Lee, J. S., Kurie, J. M., Fan, Y. H., Lippman, S. M., Lee, J. J., Ro, J. Y., Broxson, A., Yu, R., Morice, R. C., Kemp, B. L., Khuri, F. R., Walsh, G. L., Hittelman, W. N., and Hong, W. K. Clonal genetic

alterations in the lungs of current and former smokers. J. Natl. Cancer Inst., *89:* 857–862, 1997.

44. Smith, A. L., Hung, J., Walker, L., Rogers, T. E., Vuitch, F., Lee, E., and Gazdar, A. F. Extensive areas of aneuploidy are present in the respiratory epithelium of lung cancer patients. Br. J. Cancer, *73:* 203–209, 1996.

45. Franklin, W. A., Gazdar, A. F., Haney, J., Wistuba, I. I., La Rosa, F. G., Kennedy, T., Ritchey, D. M., and Miller, Y. E. Widely dispersed p53 mutation in respiratory epithelium: a novel mechanism for field carcinogenesis. J. Clin. Investig., *100:* 2133–2137, 1997.

46. Johnston, W. W., and Elson, C. E. Respiratory tract. *In:* M. Bibbo (ed.), Comprehensive Cytopathology, pp. 325–401. Philadelphia: Saunders, 1997.

47. Travis, W. D., Travis, L. B., and Devesa, S. S. Lung cancer. Cancer (Phila.), *75* (Suppl. 1): 191–202, 1995.

48. Wynder, E. L., and Muscat, J. E. The changing epidemiology of smoking and lung cancer histology. Environ. Health Perspect., *103* (Suppl. 8): 143–148, 1995.

49. Thun, M. J., Lally, C. A., Flannery, J. T., Calle, E. E., Flanders, W. D., and Heath, C. W., Jr. Cigarette smoking and changes in the histopathology of lung cancer. J. Natl. Cancer Inst., *89:* 1580–1586, 1997.

50. Lam, S., and Shibuya, H. Early diagnosis of lung cancer. Clin. Chest Med., *20:* 53–61, 1999.

51. Kennedy, T. C., Proudfoot, S., Piantadosi, S., Wu, L., Saccomanno, G., Petty, T. L., and Tockman, M. S. Efficacy of two sputum collection techniques in patients with air flow obstruction. Acta Cytol., *43:* 630–636, 1999.

52. Browman, G. P., Arnold, A., Levine, M. N., and D'Souza, T. Use of screening phase data to evaluate observer variation of sputum cytodiagnosis as an outcome measure in a chemoprevention trial. Cancer Res., *50:* 1216–1219, 1990.

53. Cantaboni, A., Pezzotta, M. G., Sironi, M., and Porcellati, M. Quality assurance in pathology: cytologic and histologic correlation. Acta Cytol., *36:* 717–721, 1992.

54. Holiday, D. B., McLarty, J. W., Farley, M. L., Mabry, L. C., Cozens, D., Roby, T., Waldron, E., Underwood, R. D., Anderson, E., and Culbreth, W. Sputum cytology within and across laboratories. A reliability study. Acta Cytol., *39:* 195–206, 1995.

55. Risse, E. K., Vooijs, G. P., and van't Hoff, M. A. Relationship between the cellular composition of sputum and the cytologic diagnosis of lung cancer. Acta Cytol., *31:* 170–176, 1987.

56. MacDougall, B., and Weinerman, B. The value of sputum cytology. J. Gen. Intern. Med., *7:* 11–12, 1992.

57. Tockman, M. S., Erozan, Y. S., Gupta, P., Piantadosi, S., Mulshine, J. L., and Ruckdeschel, J. C. The early detection of second primary lung cancers by sputum immunostaining. Chest, *106* (Suppl.): 385S–390S, 1994.

58. Tockman, M. S., Gupta, P. K., Myers, J. D., Frost, J. K., Baylin, S. B., Gold, E. B., Chase, A. M., Wilkinson, P. H., and Mulshine, J. L. Sensitive and specific monoclonal antibody recognition of human lung cancer antigen on preserved sputum cells: a new approach to early lung cancer detection. J. Clin. Oncol., *6:* 1685–1693, 1988.

59. Zhou, J., Mulshine, J. L., Unsworth, E. J., Scott, F. M., Avis, I. M., Vos, M. D., and Treston, A. M. Purification and characterization of a protein that permits early detection of lung cancer. J. Biol. Chem., *271:* 10760–10766, 1996.

60. Qiao Y-L, Tockman, M. S., Li, L., Erozan, Y. S., Yao, S. X., Barrett, M. J., Zhou, W. H., Giffen, C. A., Luo, X. C., and Taylor, P. R. A case-cohort study of an early biomarker of lung cancer in a screening cohort of Yunnan tin miners in China. Cancer Epidemiol. Biomark. Prev, *6:* 893–900, 1997.

61. Grover, F. L., and Piantadosi, S. Recurrence and survival following resection of bronchioloalveolar carcinoma of the lung: the Lung Cancer Study Group experience. Ann. Surg., *209:* 779–790, 1989.

62. Tockman, M. S. Advances in sputum analysis for screening and early detection of lung cancer. Cancer Control, *7:* 19–24, 2000.

63. Sueoka, E., Goto, Y., Sueoka, N., Kai, Y., Kozu, T., and Fujiki, H. Heterogeneous nuclear ribonucleoprotein B1 as a new marker of early detection for human lung cancers. Cancer Res., *59:* 1404–1407, 1999.

64. Mulshine, J. L. Reducing lung cancer risk. Early detection. Chest, *116:* 493S–496S, 1999.

65. Mao, L., Hruban, R. H., Boyle, J. O., Tockman, M., and Sidransky, D. Detection of oncogene mutations in sputum precedes diagnosis of lung cancer. Cancer Res., *54:* 1634–1637, 1994.

66. Ruppert, J. M., Tokino, K., and Sidransky, D. Evidence for two bladder cancer suppressor loci on chromosome 9. Cancer Res., *53:* 5093–5095, 1993.

67. Nawroz, H., van der Riet, P., Hruban, R. H., Koch, W., Ruppert, J. M., and Sidransky, D. Allelotype of head and neck squamous cell carcinoma. Cancer Res., *54:* 1152–1155, 1994.

68. Miozzo, M., Sozzi, G., Musso, K., Pilotti, S., Incarbone, M., Pastorino, U., and Pierotti, M. A. Microsatellite alterations in bronchial and sputum specimens of lung cancer patients. Cancer Res., *56:* 2285–2288, 1996.

69. Shapiro, G. I., Park, J. E., Edwards, C. D., Mao, L., Merlo, A., Sidransky, D., Ewen, M. E., and Rollins, B. J. Multiple mechanisms of $p16^{INK4A}$ inactivation in non-small cell lung cancer cell lines. Cancer Res., *55:* 6200–6209, 1995.

70. Hamada, K., Kohno, T., Kawanishi, M., Ohwada, S., and Yokota, J. Association of CDKN2A(p16)/CDKN2B9p15) alterations and homozygous chromosome arm 9p deletions in human lung carcinoma. Genes Chromosomes Cancer, *22:* 232–240, 1998.

71. Belinsky, S. A., Nikula, K. J., Palmisano, W. A., Michels, R., Saccomanno, G., Gabrielson, E., Baylin, S. B., and Herman, J. G. Aberrant methylation of $p16^{ink4a}$ is an early event in lung cancer and a potential biomarker for early diagnosis. Proc. Natl. Acad. Sci. USA, *95:* 11891–11896, 1998.

72. Esteller, M., Sanchez-Cespedes, M., Rosell, R., Sidransky, D., Baylin, S. B., and Herman, J. G. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. Cancer Res., *59:* 67–70, 1999.

73. Nieburgs, H. E. Recent progress in the interpretation of malignancy associated changes (MAC). Acta Cytol., *12:* 445–453, 1968.

74. Payne, P. W., Sebo, T. J., Doudkine, A., Garner, D., MacAulay, C., Lam, S., LeRiche, J. C., and Palcic, B. Sputum screening by quantitative microscopy: a reexamination of a portion of the National Cancer Institute Cooperative Early Lung Cancer Study. Mayo Clin. Proc., *72:* 697–704, 1997.

75. Lam, S., Palcic, B., Garner, D., Beveridge, J., MacAulay, C., LeRiche, J., and Coldman, A. Lung Cancer Control Strategy in the New Millennium. Lung Cancer, *29* (Suppl. 2): 145, 2000.

76. Ahrendt, S. A., Chow, J. T., Xu, L. H., Yang, S. C., Eisenberger, C. F., Esteller, M., Herman, J. G., Wu, L., Decker, P. A., Jen, J., and Sidransky, D. Molecular detection of tumor cells in bronchoalveolar lavage fluid from patients with early stage lung cancer. J. Natl. Cancer Inst., *91:* 332–339, 1999.

77. Gazdar, A. F., and Minna, J. D. Molecular detection of early lung cancer. J. Natl. Cancer Inst., *91:* 299–301, 1999.

78. Field, J. K., Liloglou, T., Xinarianos, G., Prime, W., Fielding, P., Walshaw, M. J., and Turnbull, L. Genetic alterations in bronchial lavage as a potential marker for individuals with a high risk of developing lung cancer. Cancer Res., *59:* 2690–2695, 1999.

79. Liloglou, T., Maloney, P., Xinarianos, G., Fear, S., and Field, J. K. Sensitivity and limitations of high throughput fluorescent microsatellite analysis for the detection of allelic imbalance. Application in lung tumors. Int. J. Oncol, *16:* 5–14, 2000.

80. Fielding, P., Turnbull, L., Prime, W., Walshaw, M., and Field, J. K. Heterogeneous nuclear ribonucleoprotein A2/B1 up-regulation in bronchial lavage specimens: a clinical marker of early lung cancer detection. Clin. Cancer Res., *5:* 4048–4052, 1999.

81. Pantel, K., Schlimok, G., Angstwurm, M., Weckermann, D., Schmaus, W., Gath, H., Passlick, B., Izbicki, J. R., and Riethmuller, G. Methodological analysis of immunocytochemical screening for dissem-

inated epithelial tumor cells in bone marrow. J. Hematother., *3:* 165–173, 1994.

82. Lambrechts, A. C., van't Veer, L. J., and Rodenhuis, S. The detection of minimal numbers of contaminating epithelial tumor cells in blood or bone marrow: use, limitations and future of RNA-based methods. Ann. Oncol., *9:* 1269–1276, 1998.

83. Steinman, C. R. Free DNA in serum and plasma from normal adults. J. Clin. Investig., *66:* 1391–1399, 1980.

84. Raptis, L., and Menard, H. A. Quantitation and characterization of plasma DNA in normals and patients with lupus erythematosus. J. Clin. Investig., *66:* 1391–1399, 1980.

85. Leon, S. A., Shapiro, B., Sklaroff, D. M., and Yaros, M. J. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res., *37:* 646–650, 1977.

86. Stroun, M., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., and Beljanski, M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology, *46:* 318–322, 1989.

87. Shapiro, B., Chakrabaty, M., Cohn, E., and Leon, S. A. Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease. Cancer (Phila.), *51:* 2116–2120, 1983.

88. Chen, X. Q., Stroun, M., Magnenat, J-L., Nicod, L. P., Kurt, A. M., Lyautey, J., Lederrey, C., and Anker, P. Microsatellite alterations in plasma DNA of small cell lung cancer patients. Nat. Med., *2:* 1033–1037, 1996.

89. Anker, P., Lefort, F., Vasioukhin, V., Lyautey, J., Lederrey, C., Chen, X. Q., Stroun, M., Mulcahy, H. E., and Farthing, M. J. K-ras mutations are found in DNA extracted from the plasma of colorectal cancer patients. Gastroenterology, *112:* 1114–1129, 1997.

90. Kopreski, M. S., Benko, F. A., Kwee, C., Leitzel, K. E., Eskander, E., Lipton, A., and Gocke, C. D. Detection of mutant K-ras DNA in plasma or serum of patients with colorectal cancer. Br. J. Cancer, *76:* 1293–1299, 1997.

91. Hibi, K., Robinson, R., Wu, L., Hamilton, S. R., Sidransky, D., and Jen, J. Molecular detection of genetic alterations in the serum of colorectal cancer patients. Cancer Res., *58:* 1405–1407, 1998.

92. Sorenson, G. D., Pribish, D. M., Valone, F. H., Memoli, V. A., Bzik, D. J., and Yao, S. L. Soluble normal and mutated DNA sequences from single copy genes in human blood. Cancer Epidemiol. Biomark. Prev., *3:* 67–71, 1994.

93. Mulcahy, H. E., Lyautey, J., Lederrey, C., qi Chen, X., Anker, P., Alstead, E. M., Ballinger, A., Farthing, M. J., and Stroun, M. A prospective study of K-*ras* mutations in the plasma of pancreatic cancer patients. Clin. Cancer Res., *4:* 271–275, 1998.

94. Vasioukhin, V., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., and Stroun, M. Point mutations of the N-ras in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. Br. J. Haematol., *86:* 774–779, 1994.

95. Woolner, L. B. Pathology of cancer detected cytologically. *In:* Atlas of Early Lung Cancer, National Cancer Institute, NIH, United States Department of Health and Human Services, pp. 107–203. Tokyo: Igaku-Shoin, 1983.

96. Kato, H., and Cortese, D. A. Early detection of lung cancer by means of hematoporphyrin derivative fluorescence and laser photoradiation. Clin. Chest Med., *6:* 237–253, 1985.

97. Profio, A. E., Doiron, D. R., and King, E. G. L. Laser fluorescence bronchoscopy for localization of occult lung tumors. Med. Phys. (NY), *6:* 532–535, 1979.

98. Kinsey, J. H., and Cortese, D. A. Endoscopic system for simultaneous visual examination and electronic detection of fluorescence. Rev. Sci. Instrum., *51:* 1403–1406, 1980.

99. Profio, A. E., Doiron, D. R., and Sarmaik, J. Fluorometer for endoscopic diagnosis of tumors. Med. Phys. (NY), *11:* 516–520, 1984.

100. Montan, S., Svanberg, K., and Svanberg, S. Multicolor imaging and contrast enhancement in cancer-tumor localization using laser-induced fluorescence in hematoporphyrin-derivative-bearing tissue. Optics Lett., *10:* 56–58, 1985.

101. Kato, H., Imaizumi, T., Aizawa, K., Iwabuchi, H., Yamamoto, H., Ikeda, N., Tsuchida, T., Tamachi, Y., Ito, T., and Hayata, Y. Photodynamic diagnosis in respiratory tract malignancy, using an excimer dye laser system. J. Photochem. Photobiol. Biol., *6:* 189–196, 1990.

102. Kato, H., and Ikeda, N. The role of fluorescence diagnosis in the early detection of high-risk bronchial lesions. J. Bronchol., *5:* 273–274, 1998.

103. Palcic, B., Lam, S., Hung, J., and MacAulay, C. Detection and localization of early lung cancer by imaging techniques. Chest, *99:* 742–743, 1991.

104. Hung, J., Lam, S., LeRiche, J. C., and Palcic, B. Autofluorescence of normal and malignant bronchial tissue. Lasers Surg. Med., *11:* 99–105, 1991.

105. Lam, S., Kennedy, T., Unger, M., Miller, Y. E., Gelmont, D., Rusch, V., Gipe, B., Howard, D., LeRiche, J. C., Coldman, A., and Gazdar, A. F. Localization of bronchial intraepithelial neoplastic lesions by fluorescence bronchoscopy. Chest, *113:* 696–702, 1998.

106. Kurie, J. M., Lee, J. S., Morice, R. C., Walsh, G. L., Khuri, F. R., Broxson, A., Ro JY. Franklin, W. A., Yu, R., and Hong, W. K. Autofluorescence bronchoscopy in the detection of squamous metaplasia and dysplasia in current and former smokers. J. Natl. Cancer Inst., *90:* 991–995, 1998.

107. Venmans, B. J., van der Linden, J. C., van Boxem, A. J., Postmus, P. E., Smit, E. F., and Sutedja, G. Early detection of pre-invasive lesions in high risk patients. A comparison of conventional fiberoptic and fluorescence bronchoscopy. J. Bronchol., *5:* 280–283, 1998.

108. Vermylen, P., Pierard, P., Roufosse, C., Bosschaerts, T., Verhest, A., Sculier, J-P., and Ninane, V. Detection of bronchial preneoplastic lesions and early lung cancer with fluorescence bronchoscopy: a study about its ambulatory feasibility under local anaesthesia. Lung Cancer (Limerick), *25:* 161–168, 1999.

109. Kennedy, T., Hirsch, F. R., Miller, Y., Prindiville, S., Bunn, P. A., Jr., and Franklin, W. A randomized study of fluorescence bronchoscopy versus white-light bronchoscopy for early detection of lung cancer in high risk patients. Lung Cancer (Limerick), *29* (Suppl. 2)*:* 244, 2000.

110. Gazdar, A. F., and Minna, J. D. Angiogenesis and the multistage development of lung cancers. Clin. Cancer Res., *6:* 1611–1612, 2000.

111. Henschke, C. I., Miettinen, O. S., Yankelevitz, D. F., Libby, D. M., and Smith, J. P. Radiographic screening for cancer: proposed paradigm for requisite research. Clin. Imaging, *18:* 6–20, 1994.

112. Miettinen, O. S. Screening for lung cancer. Radiol. Clin. N. Am., *38:* 479–486, 2000.

113. Marcus, P. M., Bergstralh, E. J., Fagerstrom, M., Willliams, D. E., Fontana, R., Taylor, W. F., and Prorok, P. C. Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up. J. Natl. Cancer Inst., *92:* 1308–1316, 2000.

114. Black, W. C. Overdiagnosis. An underrecognized cause of confusion and harm in cancer screening. J. Natl. Cancer Inst., *92:* 1280–1282, 2000.

115. The Health, and Welfare Statistics Foundation. The current issue on the screening project for elderly people by the Ministry of Health and Welfare Japan. J. Health Welfare, *42:* 37, 1995.

116. Sobue, T., Suzuki, T., and Naruke, T. The Japanese Lung-Cancer-Screening Research Group; a case-control study for evaluating lung cancer screening in Japan. Int. J. Cancer, *50:* 230–237, 1992.

117. Okamoto, N., Suzuki, T., Hasegawa, H., Goto, T., Hagiware, S., Sekimoto, M., and Kaneko, M. Evaluation of a clinic-based screening program for lung cancer with a case-control design in Kanagawa, Japan. Lung Cancer (Limerick), *25:* 77–85, 1999.

118. Muhm, J. R., Miller, W. E., Fontana, R. S., Sanderson, D. R., and Uhlenhopp, M. A. Lung cancer detected during a screening program using four-month chest radiographs. Radiology, *148:* 609–615, 1983.

119. Hayabuchi, N., Russell, W. J., and Murakami, J. Problems in radiographic detection and diagnosis of lung cancer. Acta Radiol. (CPH), *30:* 163–167, 1989.

120. Woodring, J. H. Pitfalls in the radiologic diagnosis of lung cancer. Am. J. Roentgenol., *154:* 1165–1175, 1990.

121. Greene, R. E. Missed lung nodules: lost opportunities for cancer care. Radiology, *182:* 8–9, 1992.

122. Austin, J. H. M., Romney, B. M., Goldsmith, L. S. Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesions evident in retrospect. Radiology, *182:* 115–122, 1992.

123. Schaner, E. G., Chang, A. E., Doppan, J. L., Conkle, D. M., Flye, M. W., and Rosenberg, S. A. Comparison of computed and conventional whole lung tomography in detecting pulmonary nodules. Am. J. Roentgenol., *131:* 51–54, 1978.

124. Webb, W. R. Radiologic evaluation of solitary pulmonary nodule. Am. J. Roentgenol., *154:* 701–708, 1990.

125. Kalender, W. A., Seissler, W., Klotz, E., and Vock, P. Spiral volumetric CT with single-breath-hold technique, continuous transport, and continuous scanner rotation. Radiology, *176:* 181–183, 1990.

126. Costello, P., Anderson, W. A., and Blume, D. Pulmonary nodule: evaluation with spiral volumetric CT. Radiology, *179:* 875–876, 1991.

127. Remy-Jardin, M., Giraud, F., and Marquette, C. H. Pulmonary nodules: detection with thick section spiral CT versus conventional CT. Radiology, *187:* 513–520, 1993.

128. Davis, S. CT evaluation for pulmonary metastases in patients with extrathoracic malignancy. Radiology, *180:* 1–12, 1991.

129. Kaneko, M., Eguchi, K., Ohmatsu, H., Kakinuma, R., Naruke, T., Suemasu, K., and Moriyama, N. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. Radiology, *201:* 798–802, 1996.

130. Sone, S., Takashima, S., Li, F., Yang, Z., Honda, T., Maruyama, Y., Hasegawa, M., Yamanda, T., Kubo, K., Hanamura, K., and Asakura, K. Mass screening for lung cancer with mobile spiral computed tomography scanner. Lancet, *351:* 1242–1245, 1998.

131. Henschke, C. I., McCauley, D. I., Yankelevitz, D. F., Naidich, D. P., McGuinness, G., Miettinen, O. S., Libby, D. M., Pasmantier, M. W., Koizumi, J., Altorki, N. K., and Smith, J. P. Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet, *354:* 99–105, 1999.

132. Kakinuma, R., Ohmatsu, H., Kaneko, M., Eguchi, K., Naruke, T., Nagai, K., Nishiwaki, Y., Suzuki, A., and Moriyama, N. Detection failures in spiral CT screening for lung cancer: analysis of CT findings. Radiology, *212:* 61–66, 1999.

133. Paranjpe, D. V., and Bergin, C. J. Spiral CT of the lungs: optimal technique and resolution compared with conventional CT. Am. J. Roentgenol., *162:* 561–567, 1994.

134. Seltzer, S. E., Judy, P. F., Adams, D. F., Jacobson, F. L., Stark, P., Kikinis, R., Swensson, R. G., Hooton, S., Head, B., and Feldman, U. Spiral CT of the chest: comparison of cine and film-based viewing. Radiology, *197:* 73–78, 1995.

135. Croisille, P., Souto, M., Cova, M., Wood, S., Afework, Y., Kuhlman, J. E., and Zerhouni, E. A. Pulmonary nodules: improved detection with vascular segmentation and extraction with spiral CT. Radiology, *197:* 397–401, 1995.

136. Kanazawa, K., Kawata, Y., Niki, N., Satoh, H., Ohmatsu, H., Kakinuma, R., and Kaneko, M. Computer-aided diagnosis for pulmonary nodules based on helical CT images. Comput. Med. Imaging Graph., *22:* 157–167, 1998.

137. Buckshee, N., Agnello, K., Yankelevitz, D. F., Mancuso, C., and Henschke, C. I. Smoking habits and overall satisfaction after early lung cancer screening using low-dose CT. ALA/ATS International Conference, 1999.

138. Coleman, R. E. PET in lung cancer. J. Nucl. Med., *40:* 814–820, 1999.

139. Maitra A. Wistuba, I. I., Virmani, A. K., Sakaguchi, M., Park, I., Stucky, A., Milchgrub, S., Gibbons, D., Minna, J. D., and Gazdar, A. F. Enrichment of epithelial cells for molecular studies. Nat. Med., *5:* 459–463, 1999.

140. Kopelovich, L., Henson, D. E., Gazdar, A. F., Dubb, B., Srivastava, S., Kelloff, G. J., and Greenwald, P. Surrogate anatomic/functional sites for evaluating cancer risk: an extension of the field effect. Clin. Cancer Res., *5:* 3899–3905, 1999.

141. Crowell, R. E., Gilliland, F. D., Temes, R. T., Harms, H. J., Neft, R. E., Heaphy, E., Auckley, D. H., Crooks, L. A., Jordan, S. W., Samet, J. M., Lechner, J. F., and Belinsky, S. A. Detection of trisomy 7 in nonmalignant bronchial epithelium from lung cancer patients and individuals at risk for lung cancer. Cancer Epidemiol. Biomark. Prev., *5:* 631–637, 1996.

142. Somers, V. A., van Henten, A. M., ten Velde, G. P., Arends, J. W., and Thunnissen, F. B. Additional value of K-ras point mutations in bronchial wash fluids for diagnosis of peripheral lung tumours. Eur. Respir. J., *13:* 1120–1124, 1999.

143. Yahata, N., Ohyashiki, K., Ohyashiki, J. H., Iwama, H., Hayashi, S., Ando, K., Hirano, T., Tsuchida, T., Kato, H., Shay, J. W., and Toyama, K. Telomerase activity in lung cancer cells obtained from bronchial washings. J. Natl. Cancer Inst., *90:* 684–690, 1998.

144. Wistuba, I. I., and Gazdar, A. F. Molecular abnormalities in the sequential development of lung carcinoma. *In:* S. Srivastava, D. E. Henson, and A. F. Gazdar (eds.), Molecular Pathology of Early Cancer, pp. 265–276. Amsterdam: IOS Press, 1999.

145. Yashima, K., Litzky, L. A., Kaiser, L., Rogers, T., Lam, S., Wistuba, I. I., Milchgrub, S., Srivastava, S., Piatyszek, M. A., Shay, J. W., and Gazdar, A. F. Telomerase expression in respiratory epithelium during the multistage pathogenesis of lung carcinomas. Cancer Res., *57:* 2373–2377, 1997.

146. Scott, F. M., Treston, A. M., Shaw, G. L., Avis, I., Sorenson, J., Kelly, K., Dempsey, E. C., Cantor, A. B., Tockman, M., and Mulshine, J. L. Peptide amidating activity in human bronchoalveolar lavage fluid. Lung Cancer (Limerick), *14:* 239–251, 1996.

147. Scott, F. M., Modali, R., Lehman, T. A., Seddon, M., Kelly, K., Dempsey, E. C., Wilson, V., Tockman, M. S., and Mulshine, J. L. High frequency of K-ras codon 12 mutations in bronchoalveolar lavage fluid of patients at high risk for second primary lung cancer. Clin. Cancer Res., *3:* 479–482, 1997.

148. Mills, N. E., Fishman, C. L., Rom, W. N., and Jacobson, D. R. *Ras* oncogene detection in bronchioloalveolar lavage fluid from patients with lung cancer. Lung Cancer (Limerick), *1* (Suppl.): 11, 1994.

149. Nawroz, H., Koch, W., Anker, P., Stroun, M., and Sidransky, D. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat. Med., *2:* 1035–1037, 1996.

150. Brennan, J. A., Mao, L., Hruban, R. H., Boyle, J. O., Eby, Y. J., Koch, W. M., Goodman, S. N., and Sidransky, D. Molecular assessment of histopathological staging in squamous-cell carcinoma of the head and neck. N. Engl. J. Med., *332:* 429–435, 1995.

151. Ahrendt, S. A., Yang, S. C., Wu, L., Westra, W. H., Jen, J., Califano, J. A., and Sidransky, D. Comparison of oncogene mutation detection and telomerase activity for the molecular staging of non-small cell lung cancer. Clin. Cancer Res., *3:* 1207–1214, 1997.

152. Gazdar, A. F., and Minna, J. D. NCI series of cell lines: an historical perspective. J. Cell. Biochem., *24* (Suppl.): 1–11, 1996.

153. Wistuba, I. I., Bryant, D., Behrens, C., Milchgrub, S., Virmani, A. K., Ashfaq, R., Minna, J. D., and Gazdar, A. F. Comparison of features of human lung cancer cell lines and their corresponding tumors. Clin. Cancer Res., *5:* 991–1000, 1999.

154. Amstad, P., Reddel, R. R., Pfeifer, A., Malan, S. L., Mark, G. D., and Harris, C. C. Neoplastic transformation of a human bronchial epithelial cell line by a recombinant retrovirus encoding viral Harvey ras. Mol. Carcinog., *1:* 151–160, 1988

155. Franklin, W. A., Folkvord, J. M., Varella-Garcia, M., Kennedy, T., Proudfoot, S., Cook, R., Dempsey, E. C., Helm, K., Bunn, P. A., and Miller, Y. E. Expansion of bronchial epithelial cell populations by in vitro culture of explants from dysplastic and histologically normal sites. Am. J. Respir. Cell Mol. Biol., *15:* 292–304, 1996.

156. Kelsey, K., Spitz, M., Zuo, A., and Wiencke, J. Deletion of glutathione *S*-transferase class μ and class θ genes interacts to enhance susceptibility to lung cancer in minority populations. Cancer Causes Control, 1997.

157. Wu, X., Zhao, Y., Honn, S. E., Tomlinson, G. E., Minna, J. D., Hong, W. K., and Spitz, M. R. Benzo[*a*]pyrene diol epoxide-induced 3p21.3 aberrations and genetic predisposition to lung cancer. Cancer Res., *58:* 1605–1608, 1998.

158. Virmani, A. K., Fong, K. M., Kodagoda, D., McIntire, D., Hung, J., Tonk, V., Minna, J. D., and Gazdar, A. F. Allelotyping demonstrates common and distinct patterns of chromosomal loss in human lung cancer types. Genes Chromosomes Cancer, *21:* 308–319, 1998.

159. Takahashi, T., Nau, M. M., Chiba, I., Birrer, M. J., Rosenberg, R. K., Vinocour, M., Levitt, M., Pass, H., Gazdar, A. F., and Minna, J. D. p53: a frequent target for genetic abnormalities in lung cancer. Science (Washington DC), *246:* 491–494, 1989.

160. Forgacs, E., Biesterveld, E. J., Sekido, Y., Fong, K. M., Muneer, S., Wistuba, I., Milchgrub, S., Brezinschek, R., Virmani, A., Gazdar, A. F., and Minna, J. D. Mutation analysis of the *PTEN/MMAC1* gene in lung cancer. Oncogene, *17:* 1557–1565, 1998.

161. Mitsudomi, T., Steinberg, S., Oie, H. K., Mulshine, J. L., Phelps, R., Viallet, J., Pass, H., Minna, J. D., and Gazdar, A. F. *ras* gene mutations in non-small cell lung cancers are associated with shortened survival irrespective of treatment intent. Cancer Res., *51:* 4999–5002, 1991.

162. Sozzi, G., Veronese, M. L., Negrini, M., Baffa, R., Corticelli, M. G., Inoue, H., Tornielli, S., Pilotti, S., Ohta, M., Huebner, K., and Croce, C. M. The *FHIT* gene at 3p14.2 is abnormal in lung cancer. Cell, *85:* 17–26, 1996.

163. Anbazhagan, R., Tihan, T., Bornman, D. M., Johnston, J. C., Saltz, J. H., Weigering, A., Piantadosi, S., and Gabrielson, E. Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. Cancer Res., *59:* 5119–5122, 1999.

164. Hibi, K., Liu, Q., Beaudry, G. A., Madden, S. L., Westra, W. H., Wehage, S. L., Yang, S. C., Heitmiller, R. F., Bertelsen, A. H., Sidransky, D., and Jen, J. Serial analysis of gene expression in non-small cell lung cancer. Cancer Res., *58:* 5690–5694, 1998.

165. Sozzi, G., Pastorino, U., Moiraghi, L., Tagliabue, E., Pezzella, F., Ghirelli, C., Tornielli, S., Sard, L., Huebner, K., Pierotti, M. A., Croce, C. M., and Pilotti, S. Loss of FHIT function in lung cancer and preinvasive bronchial lesions. Cancer Res., *58:* 5032–5037, 1998.

166. Tockman, M. S., and Mulshine, J. L. Sputum screening by quantitative microscopy: a new dawn for detection of lung cancer? Mayo Clin. Proc., 72: 788–790, 1997.

167. Izzo, J., and Hittelman, W. N. Characterization of multistep tumorigenesis by *in situ* hybridization. *In:* M. Andreeff and D. Pinkel (eds.), Introduction to FISH. In press, 2001.

168. Neft, R. E., Crowell, R. E, Gilliland, F. D., Murphy, M. M., Lane, J. L., Harms, H., Coons, T., Heaphy, E., Belinsky, S. A., and Lechner, J. F. Frequency of trisomy 20 in nonmalignant bronchial epithelium from lung cancer patients and cancer-free former uranium miners and smokers. Cancer Epidemiol. Biomark. Prev, *7:* 1051–1054, 1998.

169. Heppell-Parton, A. C., Nacheva, E., Carter, N. P., and Rabbitts, P. H. A combined approach of conventional and molecular cytogenetics for detailed karyotypic analysis of the small cell lung carcinoma cell line U2020. Cancer Genet. Cytogenet., *108:* 110–119, 1999.

170. Lechner, J. F., Neft, R., Gilliland, F. D., Crowell, R. E., Auckely, D. H., Temes, R. T., and Belinsky, S. A. Individuals at high risk for lung cancer have airway epithelial cells with chromosome aberrations frequently found in lung tumor cells. In Vivo *12:* 23–26, 1998.

171. Walch, A. K., Zitzelsberger, H. F., Aubele, M. M., Mattis, A. E., Bauchinger, M., Candidus, S., Prauer, H. W., Werner, M., and Hofler, H. Typical and atypical carcinoid tumors of the lung are characterized by 11q deletions as detected by comparative genomic hybridization. Am. J. Pathol., *153:* 1089–1098, 1998.

172. Petersen, I., Bujard, M., Petersen, S., Wolf, G., Goeze, A., Schwendel, A., Langreck, H., Gellert, K., Reichel, M., Just, K., du Manoir, S., Cremer, T., Dietel, M., and Ried, T. Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung. Cancer Res., *57:* 2331–2335, 1997.

173. Smith, A. L., Hung, J., Walker, L., Rogers, T. E., Vuitch, F., Lee, E., and Gazdar, A. F. Extensive areas of aneuploidy are present in the respiratory epithelium of lung cancer patients. Br. J. Cancer, *73:* 203–209, 1996.

174. MacAulay, C. E., Lam, S., Kein-Parker, H., Gazdar, A., Guillaud, M., Payne, P., LeRiche, J., Dawe, C., Band, P., and Palcic, B. Intermediate endpoint biomarkers for lung cancer chemoprevention. SPIE, *3260:* 207–211, 1998.